Ontology-driven Question Generation for Crisis Management

Teitsma, M.

2015

document version
Publisher's PDF, also known as Version of record

Ontology-driven Question Generation for Crisis Management

Marten Teitsma

Promotiecommissie:

prof.dr. B.J. Wielinga (promotor)	Vrije Universiteit Amsterdam
prof.dr. A.Th. Schreiber (promotor)	Vrije Universiteit Amsterdam
dr. J.A.C. Sandberg (copromotor)	Universiteit van Amsterdam
prof.dr. P.W. Adriaans	Universiteit van Amsterdam
prof.dr. Aldo Gangemi	Institute of Cognitive Sciences and Technologies (Rome)
prof.dr. F.A.H. van Harmelen	Vrije Universiteit Amsterdam
prof.dr.ir. B.J.A. Kröse	Universiteit van Amsterdam en Hogeschool van Amsterdam
prof.dr. M.A. Neerincx	Technische Universiteit Delft
prof.dr. P.T.J.M. Vossen	Vrije Universiteit Amsterdam

Ontology-driven Question Generation for Crisis Management
Marten Teitsma
ISBN 978-90-9029002-7
Trefwoorden: question generation, knowledge engineering, ontologies

Gedrukt door: CPI KONINKLIJKE WÖHRMANN
Afbeelding voorpagina: De grijze boom, Mondriaan, 1911

VRIJE UNIVERSITEIT

Ontology-driven Question Generation for Crisis Management

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F. A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op vrijdag 12 juni 2015 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Marten Teitsma

geboren te Leeuwarden

promotoren:	prof.dr. A.Th. Schreiber
	prof.dr. B.J. Wielinga
copromotor:	dr. J.A.C. Sandberg

Foar Fenna en
Atze Thomas

Contents

Voorwoord

Terwijl de eerste zin van een verhaal altijd als het belangrijkst wordt gezien, is het belang slechts ontleend aan het zijn van de eerste. In dit voorwoord geef ik een aantal persoonlijke ideeën over mijn dissertatie en bedank ik veel mensen zonder wie ik dit onderzoek niet succesvol zou hebben af kunnen ronden en deze dissertatie nooit het daglicht had gezien.

Toen ik het onderzoek dat uiteindelijk tot deze dissertatie leidde, begon, was in het geheel niet duidelijk welke stappen ik zou nemen. Het onderwerp zelf 'vergaren van informatie over een crisissituatie van een grote groep gewone mensen' was snel genoeg gevonden, maar hoe onderzoek te doen was een puzzel voor me. Idealistisch als ik was, dacht ik een applicatie te moeten ontwikkelen die mensen zichzelf en anderen zou helpen gedurende een crisissituatie. Het definiëren van rollen leek voor de hand te liggen, deed ik achter een bureau gedaan en evalueerde ik vervolgens met een experiment. De hoop die elke onderzoeker heeft dat zijn hypothese bevestigd wordt, zonder twijfel en duidelijk, werd snel de grond in geslagen toen dit eerste experiment was gedaan. Evaluatie van dit experiment leidde tot een herhaling met andere parameters. Maar weer werd de hypothese gefalsificeerd. Toch, na wat statistiek, was er een spoortje van bevestiging van de originele hypothese. Dit leidde tot een volgend experiment met hetzelfde materiaal maar weer andere parameters. Na dit experiment gaven mijn promotors en ik het op en concludeerden we dat de hypothese niet was te bewijzen en zochten daar redenen voor. We schreven een artikel dat na diverse pogingen werd geaccepteerd. We leerden van dit experiment dat we moesten zoeken naar situaties waarvoor wel echte-wereld data bestond en ook literatuur over was. Dit resulteerde in een succesvol experiment waarin de hypothese werd bevestigd en dat ik kon presenteren op een conferentie van enig statuur. Toen gingen we in een meer theoretische context zoeken naar de reden waarom dit laatste experiment een succes was en valideerden de uitkomsten in een experiment. De reden dat ik dit verhaal vertel, kan gevonden worden in mijn overtuiging dat wetenschap een proces is van succes en soms ook mislukking, zoals mijn dissertatie duidelijk laat zien. De weg naar wetenschappelijk bewezen stellingen is niet recht en zeker niet op voorhand bekend. Het onderzoek dat ik heb gedaan is in elk geval de ontkenning van de stelling dat wetenschappelijk onderzoek gestuurd kan worden als elk ander productieproces. Om wetenschap succesvol te maken, moet mislukking mogelijk zijn. En dat is iets wat tegenwoordig door gezaghebbende mensen vergeten lijkt.

Terugkijkend is er een lijn die mijn afstudeerscriptie voor wijsbegeerte verbindt met deze dissertatie. Mijn scriptie had als onderwerp de verschillende rechtvaardigingen voor waarnemingsovertuigingen. Ik was en ben er van overtuigd dat deze rechtvaardiging niet gevonden kan worden in een naturalistische verklaring waarin de betekenis en rechtvaardiging voor mensen van hun waarneming buiten zichzelf gevonden kan worden. Ik denk dat het meer overtuigend is, wanneer er van wordt uitgegaan dat mensen een raamwerk van betekenis ontwikkelen juist om deze waarnemingen te rechtvaardigen. Om informatie te vergaren over een situatie doe ik dan ook een beroep op dit raamwerk. Maar het is dan de vraag hoe er een overeenkomst gecreëerd kan worden tussen enerzijds menselijk redeneren of categoriseren en anderzijds hoe machines dit doen. Logica doet een poging deze kloof te overbruggen. Daarom begin ik ook met het presenteren van een formeel systeem dat een mogelijke brug vormt. Ook de semantische theorie, die ik gebruik om informatie te waarderen, is uitgekozen juist om die reden. Deze onderwerpen had ik graag verder uitgewerkt dan in deze dissertatie is gedaan. Maar er moest op een gegeven moment een einde aan komen. Het laat echter wel zien wat volgens mij belangrijk is: onderzoek dient te beginnen bij uitgangspunten om vervolgens, stapje voor stapje, op een meer praktisch niveau te komen.

Graag zou ik de Hogeschool van Amsterdam willen bedanken: die gaf mij de gelegenheid om het voor deze dissertatie noodzakelijke onderzoek te doen. Van de personen die de Hogeschool van Amsterdam gestalte geven wil ik als eerste Marjan Freriks bedanken. Zij heeft iedereen altijd gestimuleerd hun best te doen, hetgeen een houding is die past bij een educationeel instituut. Ook Cees Rijsenbrij, die het van Marjan overnam als manager, is altijd blijven geloven in de goede afloop van dit traject. In de eerste jaren heeft Jacob Brunekreeft mij begeleid. Later is deze rol overgenomen door Ben Kröse, die begon als lector toen ik al enkele jaren onderzoek deed. Ondanks de verschillende onderwerpen die we onderzochten is hij me blijven steunen, ook tijdens moeilijke momenten. Beiden bedank ik voor de vrijheid die ze me gegeven hebben mijn eigen weg te vinden.

Dan zijn er verschillende mensen mensen die op een of andere manier een rol hebben gespeeld in het traject dat ik ben gegaan. Om te beginnen wil ik Marinus Maris, toen nog verbonden aan de Universiteit van Amsterdam, bedanken. Hij heeft bij mij de interesse in het verzamelen van informatie van gewone mensen gewekt. In het eerste jaar van mijn onderzoek heeft hij mij geholpen een geschikt onderwerp van onderzoek te vinden. In deze tijd was Frans Groen mijn promotor. Toen na een jaar bleek dat mijn onderzoek een veel sterkere psychologische kant kreeg dan Frans kon begeleiden, nam Simon Jones het over, samen met Jacobijn Sandberg als co-promotor. Na een tijdje vertrok Simon naar Abu Dhabi en werd Bob Wielinga mijn promotor. Kort daarna werd de onderzoeksgroep van Bob echter opgeheven, verplaatste ik naar de Vrije Universiteit en werd ook Guus Schreiber mijn promotor, terwijl Jacobijn mijn co-promotor bleef. Graag wil ik een ieder bedanken en in het bijzonder drie hiervan, welke ik hierna zal noemen.

Als eerste wil ik graag Bob Wielinga bedanken voor het geduld dat hij met mij heeft gehad. Steeds maar weer wees hij me op zaken die ontbraken, moeilijkheden, fouten en heel veel meer. Soms elke week maar vaker om de week kwamen we bij elkaar en heeft hij me op het goede pad gehouden gedurende al die jaren. De precisie van formuleren, het vinden van de juiste statistische methode, kortom het wetenschappelijk handwerk, heb ik van Bob geleerd. Daarvoor ben ik hem zeer dankbaar.

Als tweede wil ik graag Jacobijn Sandberg bedanken voor het geloof dat ze altijd heeft gehad in de goede afloop. Jacobijn is degene die mij het langst heeft begeleid. Altijd optimistisch na weer een afwijzing van een artikel, uitleggen waarom de afwijzing toch wel juist was en mij duidelijk maken hoe het artikel te verbeteren om toch succes te hebben de volgende keer. Dit werkte, zoals te zien is aan deze dissertatie.

Als derde wil ik graag Guus Schreiber bedanken. Hij pakte me op tijdens een erg moeilijk moment gedurende mijn traject. Hij gaf me de gelegenheid mijn onderzoek voort te zetten terwijl de tijd, van te voren gegeven, eigenlijk al lang op was.

Ook wil ik mijn familie bedanken. Mijn ouders die me hebben opgevoed en de juiste vaardigheden zoals doorzettingsvermogen en zelfvertrouwen, om dit traject te doorlopen, hebben gegeven. Ze hebben in het verleden een aantal beslissingen genomen die mij de gelegenheid gaven me te ontwikkelen zoals ik heb gedaan. Beslissingen als me naar een kleuterschool sturen terwijl dit in het kleine dorp waar we woonden helemaal niet gebruikelijk was en ondanks het aanvankelijke advies van het schoolhoofd me toch naar een hogere middelbare school sturen zodat ik uitgedaagd bleef.

Mijn vrouw Inge Willemsen heeft gedurende dit traject een opmerkelijke mate van flexibiliteit getoond. Ondanks mijn eigen overtuiging betreffende de betekenis van woorden -ik ben vrij strikt betreffende betekenis- heb ik soms gebruik gemaakt van een wel heel erg creatieve semantiek om te kunnen overleven gedurende zich herhalende discussies over de planning van mijn onderzoek. Maar uiteindelijk weten we allemaal dat creatieve semantiek zand in de ogen is dat het zicht op de werkelijkheid verstoort. Slim als ze is, wist Inge dit natuurlijk en in haar zachtmoedigheid heeft ze me steeds weer toegestaan door te gaan met het doen van onderzoek en het afmaken van deze dissertatie. Deel van mijn familie zijn natuurlijk ook mijn kinderen, Fenna en Atze Thomas Teitsma, die gedurende dit traject minder aandacht kregen dan ze verdienen. Kinderen als ze zijn, pasten ze zich aan en de uitdaging voor de komende jaren wordt dan ook ze te laten wennen aan een andere ‘Heit’. Eentje die meer beschikbaar is en samen met hen avonturen beleeft.

Preface

While the first sentence of a story of some kind is often regarded the most important sentence of the whole story, it's importance is only being determined by being first. In this preface first I will give some personal thoughts about my dissertation and then thank a lot of people without whom my research would not have been successful and the dissertation never have seen completion.

When I started my research which eventually lead to this dissertation, it was not all clear which steps to take. The subject itself 'gathering information about a crisis situation from a large number of ordinary people' was established soon enough but how to do research on this subject was a puzzle for me. Being idealistic I thought that developing an application which helps people to help themselves and others during a crisis situation was what I should do. Defining roles which, to me, seemed natural and finding a formal system to infer conclusions based on answers of people was done behind a desk and evaluated during an experiment. The hope which every researcher has that his hypothesis is proven right, clear and without doubt, was soon shattered to pieces when the first experiment was done. Evaluating this experiment my promoters and I decided to redo the experiment with some parameters altered. But again the hypothesis was falsified. Still, when doing some statistics, there was a glimpse of confirmation of the original hypothesis. This lead to yet another experiment with the same material and some parameters altered. After this last experiment we gave up and concluded the hypothesis was not to be proven and stated some reasons for it. An article was produced and accepted after several attempts. Learning from this rather disappointing experience we then looked for situations about which there was real world data and literature. This resulted in a successful experiment, confirming the hypothesis and the presentation of an article at a conference of some standing. Then we looked in a more theoretical setting for the reason why the experiment was a success and validated the results with an experiment. The reason for telling this is found in my belief that science is a process of success and sometime also of failure, as my dissertation clearly shows. The road to scientific proven statements is not straight and certainly not known beforehand. The research I have done is at least the denial of the opposite statement, saying that scientific research can be managed just as any other production process. To make science successful, failure must be possible. Which is something nowadays seems to be forgotten by people who are in authority.

In hindsight there is a some story connecting both my master thesis I wrote to finish my study in philosophy and my dissertation. The subject of my thesis was the various ways to justify perceptual beliefs. I was and still am convinced that justification of perceptual beliefs can not be found in a naturalistic theory of

perception because the justification then is given from outside human thought. I think it is more convincing when one assumes a human framework of meaning which can be used to justify perceptual beliefs. To gather information about a situation I appeal to this framework. But then becomes the question how to relate on the one side human thought or categorization and on the other side how machines are doing this. Logic tries to bridge this gap and is presented in this dissertation as a starting point. The semantic theory of information is chosen for just the same reason. I would have liked to elaborate on these subjects but then I would still not have finished this dissertation. This start shows what in my opinion is important when doing scientific research: scientific research has to start at the assumptions made and gradually grows to a level of practical use.

I would like to thank the Amsterdam University of Applied Sciences for giving me the means to do research necessary for this thesis. Of the people who embody this university, first off course Marjan Freriks who has always stimulated everybody to give their best, which is the only attitude fit for an educational institute. Then Cees Rijsenbrij who took over from Marjan and as my manager kept faith in the positive outcome of this process. In the first years Jacob Brunekreeft coached me. Later this role was taken by Ben Kröse who started as lector while I was already some years doing my research. Despite the differences of subject of our respective research he kept on supporting my process even during some difficult moments. I am grateful to both for giving me the freedom to find my own path.

Then there are several people who one way or another have played a role in this process. I would like to start with Marinus Maris who interested me in information gathering from ordinary people. With him I spend my first year looking for a subject to study. In this period Frans Groen was my promotor. After a year it became clear my research was much more psychology oriented than Frans could attend and Simon Jones took over, together with Jacobijn Sandberg. When after a short while Simon moved to Abu Dhabi, Bob Wielinga became my promotor. But then the research group at the University of Amsterdam was done away with and I moved to the VU University of Amsterdam and Guus Schreiber also became my promotor, while Jacobijn kept on acting as co-promotor. I am thankful to all of them, but specially to three whom I will mention next.

First of all, I would like to thank Bob Wielinga for his patience with me. Over and over again he pointed out omissions, difficulties, failures and what more. Sometimes in weekly and more often in biweekly sessions he kept me on track for several years. The precision of formulating, finding the most appropriate statistical method, in short the scientific handwork I learnt from Bob. For which I am truly thankful.

Secondly, I would like to thank Jacobijn Sandberg for her ongoing faith in me. Jacobijn is the one who, of all people still involved, has accompanied me the longest. Always hopeful after yet another rejection of a paper, justified that decision and made clear to me how to improve it and be successful the next time. And it worked, as is shown by this dissertation.

Thirdly, I would like to thank Guus Schreiber who picked me up during a difficult moment in the process of my research. He gave me the opportunity to continue my research while the time set beforehand was consumed.

I would also like to thank my family. My parents who raised me and gave

me the right skills such as perseverance and self-reliance to fulfill this process. They also made some decisions when I was very young which gave me the opportunity to develop as I have done. Decisions such as sending me to nursery school as one of the few kids from the small village we lived in and opposing the suggestion of the headmaster of my primary school to send me to a type of school which, as eventually became clear, should not have challenged me. Not having enjoined themselves the education suitable for their intelligence, they knew the importance of that match in live.

My wife Inge who showed a remarkable flexibility of mind when concerning the end of this process. Against my own conviction about the meaning of words, I am quite strict about meaning, I sometimes had to use very creative semantic plays to survive during yet another discussion about my planning of the process. In the end we all know that creative semantics is only sand in your eyes disturbing the picture you want to see. Being a smart women, Inge knew this also off course and her gentle nature let me be able to finish my research and this dissertation. Part of my family are also my children, Fenna and Atze Thomas Teitsma, who did not receive as much attention is they are entitled to. But children, being children, adapt to the circumstances and so the challenge for the coming years will be to let them get used to another 'Heit'. Someone who is more around and together with them enjoys adventures.

Chapter 1

Introduction

In this thesis we investigate methods to elicit information about crisis situations from ordinary people. Such information elicitation can be achieved through automatic question generation applications using mobile devices. To automatically generate questions these applications require knowledge that guides the question answering process. A key question in this thesis is how this knowledge can be acquired and represented in such a way that the interaction with ordinary people proceeds in an efficient and reliable manner.

A framework is developed to formalize the description of situations and to generate questions automatically. An ontology-driven application generates questions with commonly used concepts to enhance comprehension and thereby trustworthiness of the information given by the public.

1.1 Crisis Informatics

The research which we present in this thesis is an aggregation of several fields of study such as 'Logic', 'Knowledge Engineering', 'Informatics', 'Psychology', 'Sociology' and 'Philosophy of Information'. The construction of a framework for use in crisis situations falls within the field of 'Crisis Informatics'. Crisis Informatics includes empirical study as well as socially and behaviourally conscious ICT development for crisis situations (Palen et al., 2007b).

Most research to overcome problems of information gathering and distribution during a crisis focuses on support for the emergency services in their crisis management task. Problems as how to make information levels compatible, facilitating decision making and communication between the services or experts are the main objects of research (Turoff, 2002; Pipek et al., 2012). Such a focus on the emergency services leaves out the general public as a source of information. People involved in a crisis situation can provide information about that situation and by doing so mitigate harmful effects of the crisis.

1.2 Gathering Information from Ordinary People

Research on crisis situations indicates a variety of public involvement during a crisis. People do not wait for officials helping them but are acting on their own estimate of the situation to help themselves and others. It becomes more and more accepted to regard members of the public as true 'first responders' which together are 'a powerful, self-organizing and collectively intelligent force' which will use every means possible to optimize for local conditions (Palen et al., 2010). One of those means is the increasing availability of ICT used to share information. Not only experts in crisis management convey information about the crisis at hand but also ordinary people, i.e. people with no particular knowledge of the situation they describe.

Information technology can be of use to gather information during the so called 'golden hour', i.e., the first sixty minutes after a severe trauma (van de Ven, 2006). With respect to information gathering, a focus on a centralized approach has been the usual course (Goodman and Langhelm, 2008). A centralized structure comes with a strong hierarchical reporting structure, which has been the model in use by the emergency services. Such systems tend to ignore the public as a source of information. Our intended system is (partly) decentralized, i.e., the application runs on a mobile phone and makes use of ordinary people who happen to be in the disaster area. Until now grassroots participation of citizens during a disaster as a valuable contribution to information gathering, has not been fully appreciated by emergency services and other formally involved parties (Palen et al., 2007a). Due to this lack of appreciation, efforts to develop a technological platform to enable such participation are limited.

With the advent and combination of the internet and mobile devices such as mobile phones a new way of conveying information emerges which contrary to traditional ways of gathering and distributing information is seemingly instantly available for everyone having access to the WWW. Smartphones provide a new way of communicating on a large scale. Applications to distribute or gather information and additional sensors can be developed which make the mobile device much more fertile. Gathering information from sensors with these applications can be done opportunistically or participatory.

When only a small effort from the user is asked it is called opportunistic sensing. It is opportunistic because the movement of the sensor just follows the movement of the carrier of the device (Payton and Julien, 2010; Lane et al., 2008; Campbell et al., 2008; Zhou et al., 2012). An example is the sensing of time and location of a public transport vehicle by a mobile device which is used to inform other people about the arrival time. Opportunistic sensing lends itself to large scale deployment and application diversity because the user of the mobile device is not disrupted in his activities (Lane et al., 2008).

When the human who carries the mobile device is involved in the sensing process it is called participatory sensing. A popular means of communicating using mobile devices is microblogging. Twitter, which is one of the most used microblog platforms, was founded in 2006 and counted more than 500 million users in June 2012 (Semiocast, 2012). It allows users to send, react on and resend messages to other people known as followers. These messages are called tweets and have a maximum length of 140 characters. A growing number of people uses microblogging to disseminate information about a crisis situation. But the use of microblogs has several known disadvantages: absence of geolocation information and lack of trustworthiness (Vieweg, 2012; Terpstra et al., 2012; Hughes and Palen, 2009; Starbird and Palen, 2010). These disadvantages are the reason why organizations in the field of crisis management are reluctant to use microblogs as a source of information about a crisis at hand (Tapia et al., 2011; Dillingham et al., 2011). Furthermore, microblogging leaves the choice of subjects to the user.

The human observer can be stimulated to participate more than in microblogging by directing him at charting a particular situation. The responses can become much more versatile and adaptive. In fact, the sensing process becomes an information gathering process in which the requester can lead the human 'sensor' to provide the kind of information which is required.

1.3 A Participatory Sensing Application for Situation Awareness

We have constructed a framework which by gathering information from ordinary people determines the specific situation they are involved in. This will enhance situation awareness which is defined by Endsley (Endsley, 2000) as 'the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future'. The application we developed makes it possible to determine situations by using answers to questions which are automatically generated.

The traditional model of observation describes an observation as a one-directional flow of information (Hall and Jordan, 2010). In this model a subject receives different types of signals from its environment. After these signals are sensed by a human they are transported by neurons and ready for perception. During perception a subset of the signals is being transformed into a cognitive entity (thought or feeling). A cognitive entity is translated into language which is articulated in a text or spoken utterance.

In our daily life it is normal to experience a conscious observation of a specific phenomenon when someone turns his attention to it. The control of attention by asking questions is what we are after by directing users attention to specific objects or events in his surroundings. In fig. 1.1 this is represented by the bidirectional flow of information. We have adapted the traditional model of Hall and Jordan (Hall and Jordan, 2010) to incorporate the possibility of focusing on specific parts of the environment. The model in Figure 1.1 shows that cognition has influence on perception. This influence is realized by the attention given to specific parts of the total amount of signals being received. Such a relation gives us the opportunity to direct the gathering of information. We ask the user to attend to specific aspects of the situation by asking questions.

1.4 Problem Statement and Research Questions

In the context of a crisis situation a system for information gathering has to gather as much information as possible in a time frame which is as short as

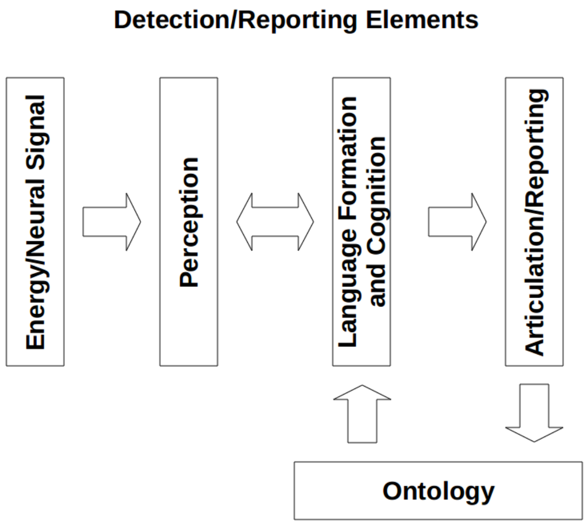


Figure 1.1: Conceptual framework for understanding intentional observations adapted from Hall and Jordan (Hall and Jordan, 2010).

possible. In our system information gathering is done by automatically generating questions which are put to the users of the system. The questions ensure a sufficient level of detail. People are employed as an omni sensor to gather heterogeneous information. The answers are used to reason about the situation people are in. A mobile device such as a mobile telephone is used as a platform.

When humans are part of the information gathering process it has to be taken into account that human reasoning differs from formal reasoning as applied by computers. A framework to reason about situations has to be constructed in such a way that this difference is overcome in a sufficient way.

The general research question is:

> Is it possible to automatically generate questions which can be answered in such a way by ordinary people that the nature of a crisis situation can be determined in a reliable way?

This general question is decomposed into several specific research questions:

RQ I. What is an appropriate framework for the formalization of crisis situations?

Situations come in all sorts and are difficult to define. We need a formalization of situations which enables us to represent real-world situations and reason about them. Such a framework will make it possible to combine pieces of information and draw conclusions.

RQ II. Given a model to formalize situations, how can we use this model to generate relevant questions?

With the help of the framework we have to generate questions and then use the answers to reach conclusions. For this we need a formalization of questions and answers. When several questions are possible, algorithms have to be created to automatically generate questions.

RQ III. What methods are suitable for creating a domain specific ontology for crisis situations?

Several methods to generate an ontology of concepts can be thought of. It has to be researched which method is best suited for the creation of such an ontology in terms of efficiency and completeness.

RQ IV. What methods can be used to evaluate the 'fit' of domain specific ontologies with human categorization?

The compliance of these ontologies with human categorization has to be evaluated. A set of metrics to evaluate the ontologies in this respect has to be constructed.

RQ V. What is the reliability of answers from ordinary people to automatically generated questions about situations?

To verify whether the framework and formalization of situations generates questions which people understand and answer in a reliable way, experiments have to be conducted. These experiments have to simulate a real-life situation where an ordinary user of the application answers questions.

1.5 Approach

Our approach to answer these research questions uses a variety of methods. First we study literature and combine several theories in a framework which can be used to formalize situations and generate questions. This framework enables us to represent knowledge with an ontology about situations in the real world.

Secondly, due to specific characteristics of ontologies several strategies to automatically generate questions have to be developed. These strategies are focussed on gathering maximal information in minimal time given the characteristics of the ontology.

The third method consists of the creation of ontologies to represent specific situations. Several ways to construct ontologies will be evaluated. The evaluation specifies different aspects of the ontology, one of which is the 'fit' with human categorization.

A fourth method we use is the creation of a prototype of an application. The prototype uses a mobile phone, the framework and several strategies to automatically generate questions.

The last method we apply is to conduct user experiments. During these experiments participants will answer the automatically generated questions. The results of these experiments are an indication of the compliance of our model with human categorization.

1.6 Outline by Chapter

This thesis starts in chapter 2 en 3 with a framework to automatically generate questions. Using theories proposed by others, we developed an ontology and several strategies to ask questions. An application to be used during experiments is implemented by the author of this thesis. In the subsequent chapters we present experiments that are published and conducted by the first author.

In chapter 2 a framework for a formalization of situations is presented. This framework gives the opportunity to abstract from specific situations and uses an ontology as representation of a real-world situation.

In chapter 3 several strategies for automatically generating questions are presented. The structure of the domain knowledge and how this is represented in an ontology restricts the strategy to use for automatically generating questions. Furthermore, an application which uses the techniques described in chapter 2 is presented. The application uses a mobile device as platform. A server is used for the distribution of the application and specific ontologies and gathers the answers of the users.

In chapter 4 an experiment with a representation created by experts is presented. This experiment shows that conclusions drawn from such an ontology are not reliable indicators for the specific state a human is in.

In chapter 5 an experiment with a pragmatic representation, constructed with concepts used by emergency workers, is presented. This experiment shows that answers to questions generated from a pragmatic-based ontology are reliable to a high degree.

In chapter 6 we investigate three methods to create an ontology. The first method uses the vocabulary employed by the Amsterdam fire department. The second method uses existing ontologies created by experts. The third method elicits concepts and their attributes from ordinary people. Metrics are used to evaluate the ontologies on four aspects: the structure of the ontology, efficiency of construction, completeness and 'fit' with human categorization.

In chapter 7 an experiment is presented using the framework and participants to evaluate the ontologies. During this experiment all three ontologies which were created in chapter 6 are used to generate questions. Analysis of the answers of the participants and the time it took for them to answer, show that characteristics of the ontologies relate to their suitability for the task of generating questions to determine a situation, as presented in chapter 6.

In chapter 8 a summary of the answers to the research questions is given. These answers show that only to a certain extent questions can be generated that ordinary people can comprehend and answer in a reliable way. A discussion about the results and a presentation of future work complete this thesis.

Chapter 2

A Formal Framework

To automatically generate questions about a situation we need a formalization of situations. Situation Theory offers a framework to formalize situations and information about what is going on in the world. Within this framework abstraction is created with parameters and types to refer to possible informational entities and situations. To represent knowledge about situations in general and within specific domains we use Situation Theory Ontology. After a revision to make this ontology compliant with knowledge engineering principles we have a framework to automatically generate questions to determine a situation.

2.1 Introduction

To formalize situations we use Situation Theory, which was initiated by Barwise and Perry (Barwise and Perry, 1983) to create a mathematical framework for the interpretation of natural languages. Subsequent research on this framework resulted in a theory useful for application in various fields of information technology. According to van Benthem, Situation Theory is better suited than other formal systems to represent information because it has a richer structure and thus is much more fine-grained (van Benthem and Martinez, 2008). A well known example of an alternative formalism to represent situations is 'Situation Calculus' (McCarthy, 1963; McCarthy and Hayes, 1968; Levesque et al., 1998). Situation Calculus however provides a much less rich structure than Situation Theory. A more recent formalism in which situations could be represented is the 'Simple Event Model' (Van Hage et al., 2011; van Hage and Ceolin, 2013). The Simple Event Model is well suited to represent dynamic situations, but it lacks structures that Situation Theory provides, such as typing and abstraction.

Situations, formalized in Situation Theory, are supported by infons which are informational entities representing a corresponding structure in the real world. Constructs offered by Situation Theory are suitable for computation. To discuss how information can be represented and combined we use terminology as presented by Devlin to describe Situation Theory (Devlin, 1991).

A mapping of Situation Theory to an ontology represented in the Web Ontology Language OWL is not straightforward. An example of such a mapping is Situation Theory Ontology (STO) (Kokar et al., 2009). We reconsidered some of the design decisions by Kokar et al. and created a revised version: Situation

Theory Ontology Revised (STOR). With this ontology we have a framework to represent real-world situations and a reasoning mechanism to draw conclusions concerning the situation about which a user is reporting.

Situation Theory is elaborated in Section 2.2. Mapping of Situation Theory to OWL is presented in 2.3. A discussion about choices made concludes this chapter.

2.2 Situation Theory

First the concept 'situation' is elaborated and then the fundamental informational entity called 'infon' are elaborated. Abstraction of infons, types and rules which generate information are discussed subsequently. In this section about Situation Theory the mathematical theory of information as developed by Devlin (Devlin, 1991) is followed.

2.2.1 Situations

A situation is a limited part of the world in which various, abstract or physical, entities stand in relation to each other. Situations are ubiquitous in our world. We are always in some situation or other. The use of 'situation' is seemingly unlimited and can stretch from being involved in an unpleasant situation after a faux pas to 'the situation in the Middle East'. Situations are also fluid and vague. They evolve into other situations and the boundaries of a situation are not always clear. Despite the vagueness surrounding situations, humans are good at recognizing a specific situation when needed. We know what is important and what not, have knowledge about consequences of specific facts and handle such fluidness and vagueness well. When something complex such as a flooding happens, it becomes the situation which urges us to take proper action without knowing all details. Being aware of such a situation means, according to Endsley (see Section 1.3), one is aware of specific elements in an environment, has a grasp of its meaning and how this constellation will evolve.

A situation cannot be completely described because it is always possible to extend a description with additional information about that situation. When, for example, we talk about a football game everybody concerned knows what is meant by 'the football game'. A talk about a football game would not be interesting if no new information was exchanged. When this *new* information is added, the situation of the football game does not necessarily become a different situation, it is just extended in a natural way.

An extensional definition of a situation, i.e. a definition with reference to a specific set of real objects, will always be incomplete about a given situation and can only represent that situation in an abstract and approximate way. The explanation referring to what is part of the situation and what not, is always in need of something which is not part of that situation, i.e. an idea which defines the situation and tells what is part of the situation and what is not. Such an idea can change even dramatically during a discussion about a situation given new information and consequently presenting the whole situation in a new perspective.

What happens and what is said always will happen or be said in some situation or another. When, for example, something is said about a particular object this is said *within the context* of the situation the speaker is in *and* what is said, i.e. the information about the object talked about, has to be understood as being part of the situation that object is in. Within Situation Theory a situation supports specific informational entities.

The informational entities which are supported by the situation, together give structure to the situation, i.e. situations consist of objects which have a specific relation to each other and confirm to specific rules. In contrast with objects which are externally defined, a situation is defined by its internal structure. An example is a game of chess played by two people according to specific rules. A set of minimum requirements make a situation a specific situation such as a game of chess. Additional information does not necessarily makes it another situation. When it is told the game is played in Amsterdam or one of the player is a grandmaster, such information is not defining the situation 'a game of chess' but it is non-defining attributing information.

Within Situation Theory different types of situations are being distinguished because of their role in the description of the world. Communicating about a situation is done within a specific situation, i.e. the *utterance situation*. The situation which has the attention and is talked about is called the *focal situation*. This is the situation which an utterance is about. Often a reference to another situation is made when doing an utterance. When referring to 'a man seen before' one is referring to a *resource situation*, i.e. the situation in which one has seen that man.

The difference between a true description and the situation itself is, in Situation Theory, the difference between what is called the *actual situation* and the *real situation*. The actual situation is an abstraction from the real world because it is a description of the real world and can not (in principle) include everything there is. Moreover, an actual situation is limited and modelled by someone who wants to talk about it. When a description refers to more than one possibly actual situations it is called an *abstract situation*.

Because a situation refers to only a part of the world, this could even mean that the same statement about something out there is true in one situation and false in another situation. A situation as a representation of a part of the world is in principle incomplete: when something is not stated it does not mean it is not a fact. Situation Theory is founded on an open world semantics which is not closed under negation, i.e. when a description is not supported by a situation it does not mean the description is false, it just means it is not part of the description of the situation.

2.2.2 Infons

In Situation Theory the smallest entity of information is called an infon which is formally described as a tuple of the form:

$$<< R, a_1, ..., a_n, 0/1 >> \tag{2.1}$$

where R is a n-place relation, and $a_1, ..., a_n$ represent objects appropriate for R. Each relation has its own specific structure, i.e. the objects appropriate for R have an assigned place in the infon. The last item is the polarity of the infon. When it is 1, the infon is a true description of a particular real situation and when it is 0, it means the infon is a false description of that real situation.

Contrary to classical logic, this does not mean the infon is always true or false in every situation. Infons are true or false within a specific actual situation referring to a specific real situation. When the real situation alters into another real situation the polarity of the infon can alter. Such true descriptions of a real situation, or a part thereof, are called 'states of affairs' and used to describe (a part of) the world.

To describe a real situation we most often also refer to a location where and a time when something took place. When these references are added to an infon their position is also specified as in the example of an infon describing that a street is flooded, which has already been determined as having a location ('Oude Tonge') and time ('February 1st, 1953'):

$$Flood_{Goeree\ Overflakkee\ 1953} \models \begin{array}{l} << flooded, street, Oude\ Tonge, \\ February\ 1st\ 1953, 1 >> \end{array} \tag{2.2}$$

where $\models$ should be read as 'supports' instead of the traditional 'makes true'. Situations are only a part of the world and in that part of the world this infon is true. This infon refers to a street which is flooded in a place called Oude Tonge at February 1st, 1953 and is supported by the situation describing the flooding of a part of the Netherlands (Goeree Overflakkee) in 1953.

A situation is represented by a set of infons which together are the description of a real situation. This is the minimum number of facts defining the situation. When a set of infons, which is sufficient to define the situation, corresponds to the real situation the description of the situation becomes actual. For example, a flooding is defined by a set of infons stating that streets are flooded. Because a flooding is a spatial phenomenon, the flooding of streets at several places at the same time is required to define a flooding. Supposing the island of Goerree Overflakkee only has two villages, Oude Tonge and Nieuwe Tonge, a situation describing the flooding of Goerree Overflakkee at February 1st, 1953, would then encompass the following propositions:

$$Flood_{Goeree\ Overflakkee\ 1953} \models \begin{array}{l} << flooded, street, Oude\ Tonge, \\ February\ 1st\ 1953, 1 >> \\ \wedge \\ << flooded, street, Nieuwe\ Tonge, \\ February\ 1st\ 1953, 1 >> \end{array} \tag{2.3}$$

When these *supported* infons correspond to what is happening in the world, the situation describing a flooding in Goeree-Overflakkee in 1953 is actual. Other *attributing* infons can be added when expedient but will not be part of the definition of a situation . During a flooding not only the streets are flooded. A lot of other things happen. Such information can be added to the situation without denying the occurrence of that situation.

2.2.3 Types and Parametric Infons

When infons have parameters, they are called parametric infons. A parametric infon is not referring to an actual situation but to possible situations, i.e. it is not clear which specific referent is meant when a parameter is used in an infon. A parametric infon is anchored by a function when that function assigns a referent for every parameter that occurs free in the infon and the infon is supported by a situation. An example of a parametric infon is:

$$\sigma = << flooded, street, \dot{l}, \dot{t}, \dot{p} >> \tag{2.4}$$

where $\dot{l}$ is a parameter for a location, for example $Oude\ Tonge$, $\dot{t}$ is a parameter for time, for example $February\ 1st\ 1953$ and $\dot{p}$ is a parameter for the polarity, i.e. 0 or 1. It is important to note that such an infon, supported by the situation, does not make a situation actual. The parametric infon σ states that somewhere a street is flooded but not where. When focused on the flooding of *Goeree Overflakkee* while $\dot{l}$ is referring to a place in an adjacent region, for example *Schouwen Duiveland*, this infon is not supported by the situation $Flood_{Goeree\ Overflakkee\ 1953}$ but by another situation. To yield information in this example, $\dot{l}$ first has to be anchored to some specific object. In this example only the parameter $\dot{l}$ is anchored by the function f to $Oude\ Tonge$ when the infon is supported by the situation $Flood_{Goeree\ Overflakkee\ 1953}$.

Infons can have parameters of a given type. Parameters can be of the basic types:

- TIM: the type of a temporal location. Refers to a specific time or time frame. For example, 2.13 PM.
- LOC: the type of a spatial location. Refers to a place such as a city, region or something else which has a location. For example, Utrecht in the Netherlands.
- IND: the type of an individual. Refers to an object which is individuated by someone. For example, the laptop computer I am writing on, also known as 'Boniface'.
- REL_n: the type of an n-place relation. For example, observing, which is a two place relation: somebody observes something.
- SIT: the type of a situation. For example, $Flood_{Goeree\ Overflakkee\ 1953}$ is a situation. The type of situations referred to, are already identified.
- INF: the type of an infon. Refers to (sub-)types which can be distinguished such as elementary infons, 'parametric infons' (infons with a parameter) and 'compound infons' (a set of infons related by conjunction and disjunction operators).
- TYP: the type of a type. Every type T is a subtype of TYP: $<< of-type, T, TYP, 1 >>$.
- PAR: the type of a parameter.
- POL: the type of a polarity (0 and 1).

For each basic type T other than PAR, there is an infinite collection of T_1, T_2, T_3, … of basic parameters, used to denote arbitrary objects of type T, i.e. such a basic parameter refers to a subtype of type T. The type of a type (TYP) defines the basic types and the type of a parameter tells what specific type of parameter is meant, for example restricted or not. We use $\dot{l}, \dot{t}, \dot{a}, \dot{s}$, etc. to denote parameters of the types LOC, TIM, IND, SIT, respectively.

Basic types such as IND, TIM can have a subtype which is a reference to a subset of the objects referenced by the basic type. This subset is restricted by conditions to objects and is determined over some initial situation. An object is always situated within a context which is described by a specific set of infons. Reference to such an object includes the context in which it is situated. Without such a situation these types would be meaningless because referring to all possible situations or another situation would refer to different (restricted) types. When, for example, someone mentions ‘a ship’ it is usually understood as a reference to ‘an object floating on the water’ and not ‘a space ship’. When in the future travel in space becomes actual, reference to the context will be necessary to avoid misunderstanding about what kind of ship is meant.

With (parameterized) infons singular facts about the world are communicated. More complex forms of information need to be handled by types which abstract over sets of infons. An example of such a set of infons is a situation $Flood_{Goeree\ Overflakkee\ 1953}$ as described in equation 2.3. The conditions which restrict parameters and object-types will be defined according to the domain in which situations are described. Each domain has its own language to convey information. This domain-specific language is an implicit definition of concepts used in the domain. The difficulty is then of course to explicate these definitions and describe them in such a way that they can be made formal and exchangeable with other domains.

A parameter can be restricted by conditions. Such restrictions exclude otherwise possible referents, i.e. not the whole set of possible referents is allowed as referent for the parameter but a subset thereof. A condition on a parameter is an infon which denotes a set of referents that satisfy the condition. Let $\dot{r}$ be a parameter, then

$$\dot{r} = v \upharpoonright C \tag{2.5}$$

which states that $\dot{r}$ denotes the referents within the set v which satisfy the condition C. Such a condition restricts the possible referents within the set of referents v.

The condition C is stated in one or more infons which have to be confirmed. An example of such a restricted parameter is:

$$\dot{r}_1 = LOC_1 \upharpoonright << partOf, Goeree\ Overflakkee, LOC_1, 1 >> \tag{2.6}$$

where a function f anchors $\dot{r}_1$, i.e. a restricted parameter that in this infon refers to a referent of the type LOC, in a situation $Flood_{Goeree\ Overflakkee\ 1953}$. When the referred object is *Oude Tonge* or *Nieuwe Tonge* the condition is satisfied. When, for example, the referred object is $Nieuwekerk$ the condition, i.e. $<< partOf, Goeree\ Overflakkee, LOC_1, 1 >>$, is not satisfied because *Nieuwekerk* is not part of *Goeree Overflakkee* but is part of another island: *Schouwen Duiveland.* In such a case the infon is not supported by the situation $Flood_{Goeree\ Overflakkee\ 1953}$, but the infon is instead supported by another situation $Flood_{Schouwen\ Duiveland\ 1953}$ because:

$$\dot{r}_2 = LOC_2 \upharpoonright << partOf, Schouwen\ Duiveland, LOC_2, 1 >> \tag{2.7}$$

and $Nieuwekerk$ is part of *Schouwen Duiveland.*

Situation Theory makes a distinction between object and situation types. Object types represent a set of all objects that make a certain set of parameterized infons true. Situation types represent the set of all situations that support a given set of infons. An important difference between object and situation types is that object types are determined by external sources, e.g. an entity is of type dog if a person perceives it as a dog and situation types are determined by their internal structure, i.e. a situation is a car accident if it involves one or more vehicles.

2.2.4 Constraints

Information in Situation Theory, is captured by the confirmation or denial of the relation between objects or situations. When, for example, someone is determining an object on fire as a ship on fire, this is information gathered by our system. When this ship, after further questioning, is determined as a cruise ship, this is also information deemed valuable for determining the situation.

Information can also be gathered by looking at the relations of causality between situations. When we have knowledge of the rule that A always is followed by B, C or D, then knowing A can make us look for distinctive infons supported by B, C or D to see which of these subsequent situations is actual.

To generate such information with this apparatus we need constraints. Constraints are relations between types of situations which represent (natural) laws, conventions and other kinds of regularities. When there is the fact of smoke somewhere, it is because of the constraint ‘fire produces smoke’ that we have a clue there is fire. In Situation Theory different types of constraints are distinguished. Nomic constraints are of the kind which correspond to some natural law, e.g. ‘fire produces smoke’. Necessary constraints are reflexive about a situation and tell more about the situation, e.g. ‘kissing means touching’. Conventional constraints refer to social laws or rules, e.g. ‘the ringing bell means class is over’.

2.3 Mapping Situation Theory to OWL

An ontology is used to construct a representation of (a part of) the world. An ontology is an explicit specification of a conceptualization (Gruber, 1995). With an ontology it is possible to reason about the concepts and their properties.

One of the most popular languages to represent ontologies is OWL. To represent Situation Theory, classes must also be instances of classes which can not be stated in OWL-DL (Baader and Nutt, 2007). So, we will map the Situation Theory formalism to OWL-Full (Dean and Schreiber, 2004). Furthermore, because OWL is commonly used, a lot of tools are available. To design ontologies we used Protégé-OWL (Knublauch et al., 2005).

First we describe in Section 2.3.1 a mapping of Situation Theory to OWL by Kokar et al. Then we reconsider some of the design decisions of Kokar et al. and present a revised version (STOR) in Section 2.3.2.

2.3.1 Situation Theory Ontology

Situation Theory is used by Kokar (Kokar et al., 2009) to create an ontology which goes under the name of Situation Theory Ontology (STO). According to Kokar et. al. the two basic elements of Situation Theory are objects and types. The construction of STO is then based on the idea that an ontology of Situation

Theory should have two meta-levels representing objects, i.e. things in the world and types which are abstractions. Furthermore, they interpret a class as a set of instances associated with this class. The first meta-level is representing objects which all are subordinate to the class *Object*. An example of such a subordinate concept is *Individual*. The second meta-level is representing types such as *IND* and *RELn* (see Section 2.2.3) as subordinate concepts of the concept *TYP*. Instances of this class are classes themselves, e.g. an instance of *IND* is the class *Individual*. In STO each kind of object has two associated classes: a class which is a set of instances of the given class and a class which is an instance of a subtype of TYP.

In Figure 2.1 a representation of STO is shown. Not all the types as presented in Situation Theory (see Section 2.2.3) are direct subordinate concepts of *TYP* and in STO some new types are defined. The types *TIM* (for timestamps) and *LOC* (for locations) are not direct subordinate concepts of *TYP* but are subordinate to *ATTR* (for attributes) which also has other subordinate concepts representing attributes of interest such as *Velocity*. In STO the type *POL* (for polarity of infons) is made subordinate to the new type *VAL* which also is superordinate concept of values such as *5* or *10*. The type *PAR* (for parameter) is omitted from STO because OWL has properties which can be defined as having a domain and range to simulate restricted parameters. A new type is *DIM* (for dimensions) which captures information about systems of units to express particular values such as *km/h* or *m* (for meter). The last new type is *RUL* which represent the rules to use in the domain which are used for reasoning.

The central concept of STO is *Situation* which has three special subclasses: *UtteranceSituation*, *FocalSituation* and *ResourceSituation* (see Section 2.2). In STO utterances are interpreted as queries of the user of the system. An utterance or its context *UtteranceSituation* is then a perspective which defines what is relevant for a specific situation. Instances of *FocalSituation* refer to specific situations which can become actual. *FocalSituation* has therefore various subclasses which restrict the general and abstract class further.

In STO each *Situation* has 6 properties. One or more *ElementaryInfons* are *supportedInfons* and are related by the property *supports*. Each *Situation* also has a *relevantRelation* and a *focalRelation*. A *relevantRelation* is relevant to a situation. An *Individual* when participating in the situation is a *relevantIndividual* and when the situation is a *UtteranceSituation* at least one *focalIndividual* is related to that situation. Finally, a *Situation* can have *Attributes*.

The class *Individual* is used for individuals, i.e. individuated objects and persons. When STO is used, a concept tree capturing the domain will be added, just as situations specific for a domain are added to *FocalSituation*. An *Individual* has an *Attribute* and an *Attribute* can have a *Dimensionality* and a *Value*. A *Value* is a superordinate concept for *Polarity* which itself is a property of *ElementaryInfon*.

2.3.2 STO Revised (STOR)

To develop an ontology we distinguish several types of ontologies: domain, generic and representation ontologies (van Heijst, 1995). The domain ontology expresses conceptualizations specific for a particular domain. A generic ontology contains concepts which are considered to be generic across several domains. A representation ontology consists of concepts which are the primitives for the

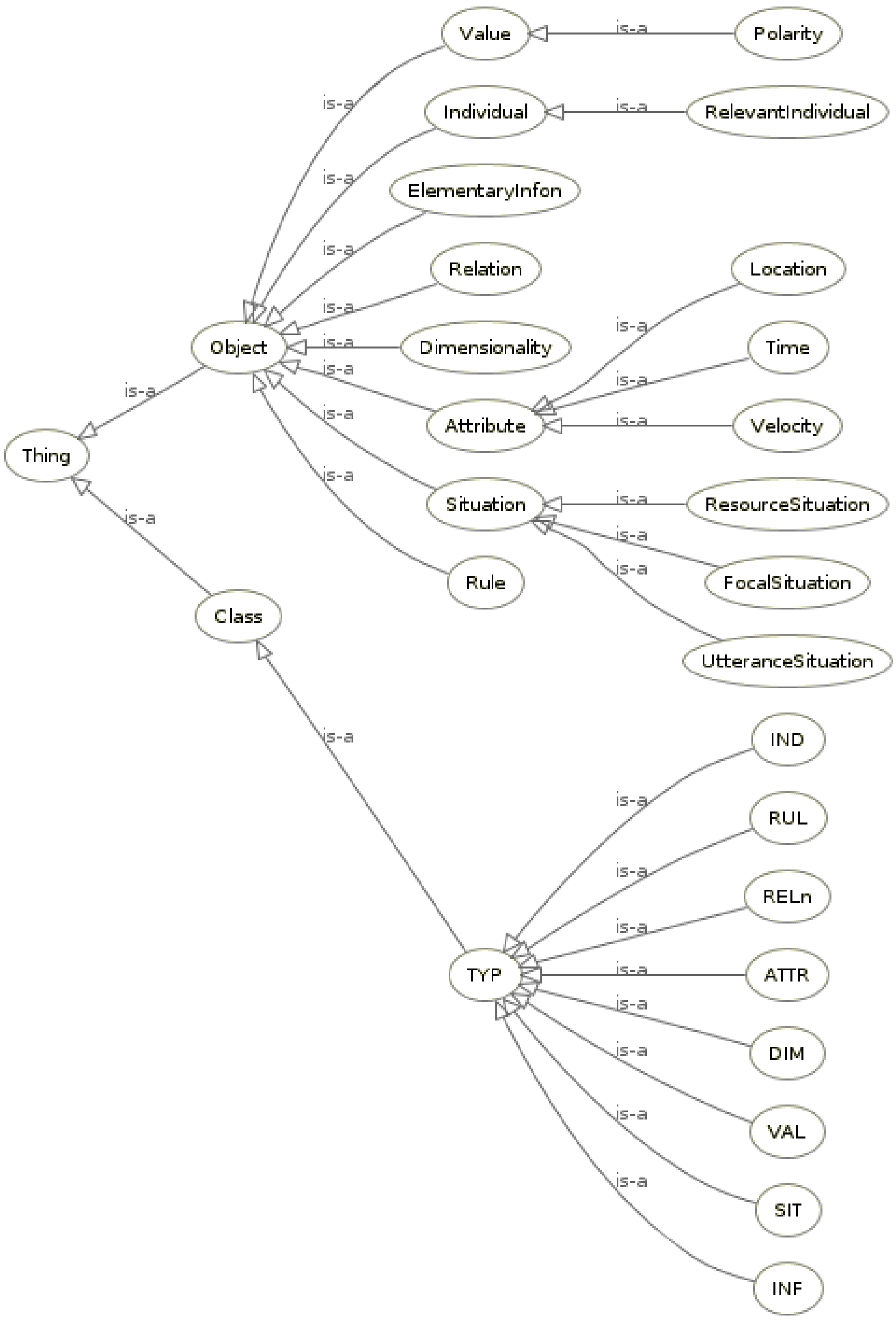


Figure 2.1: Main classes and properties of Situation Theory Ontology (adapted from (Kokar et al., 2009)).

formalization of the concepts as described in the generic and domain ontologies.

The basic types (see Section 2.2.3) should be part of the representation ontology. All the basic types are subconcepts of the type *TYP*. As can be seen, when comparing Figure 2.1 and Figure 2.2, we adjusted the types which can be

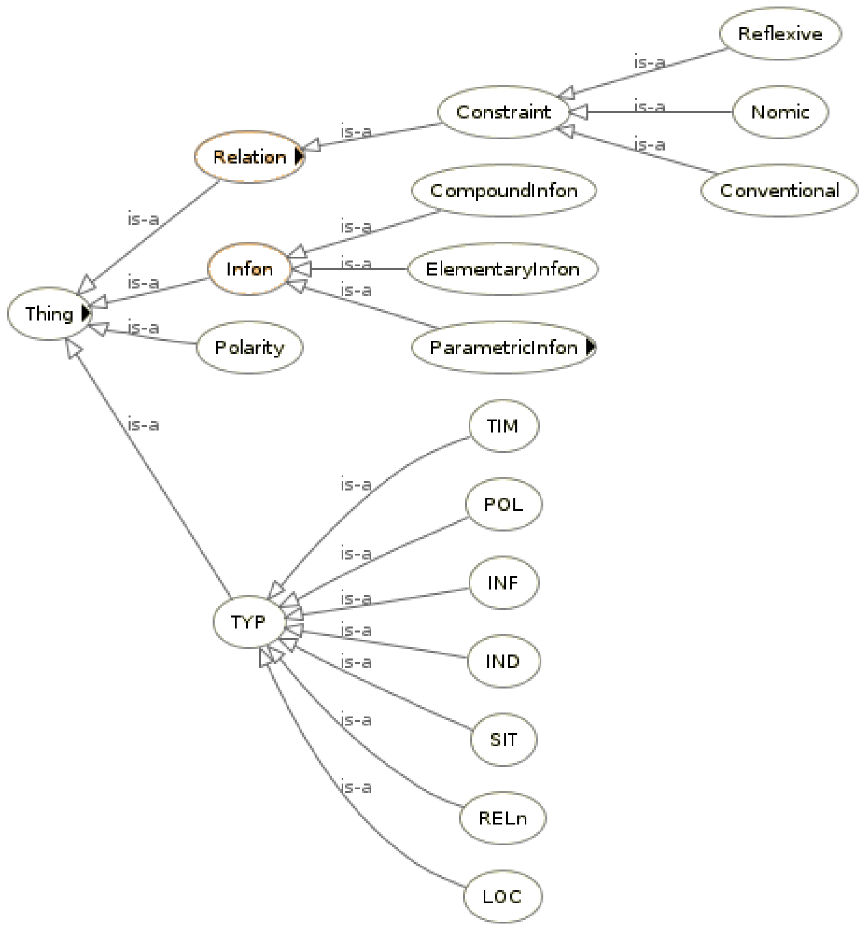


Figure 2.2: The representational part of STOR.

used in the ontology. In STO the concept *Attribute* primarily is a superclass of the concepts *Location* and *Time* which are used to represent respectively instances of location and time. But it is also used as a superclass of other attributes such as *velocity*. Attributes are, in our view, compound infons or concepts of the type *IND*. For example, *velocity* is a compound infon combining two infons referring both to the same instance of *IND* but with a distinct pair of instances of *LOC* and *TIM*. Thus *ATTR* is rejected as type in STOR and *LOC* and *TIM* are basic types (as in Situation Theory). The same can be said of the types *DIM* and *VAL*. In STO *ATTR* is also used to add properties to a *Situation* while in STOR this is done by the property *hasAttributingInfon*.

Relation and *Infon* are concepts fundamental to Situation Theory and are part of the representation ontology because these concepts have a specific formal representation (see Section 2.2.2). The concept *Infon* is a superclass of *ElementaryInfon*, *CompoundInfon* and *ParametricInfon*. The concept *ParametricInfon* is used for the generation of questions: we ask for the actual anchoring of a parameter, i.e. each parameter gives rise to a question. Instead of *Rule* as in STO (see Fig. 2.1) we use the concept *Constraint*, as subclass of *Relation*, with its subconcepts *Nomic*, *Conventional* and *Reflexive*. In OWL the constraints are represented using the type *RELn*, i.e. relation, to define a relation between situations. *Situation* is superordinate to domain-specific situations and as such part of the generic ontology. The representational part of the ontology is shown in Fig. 2.2

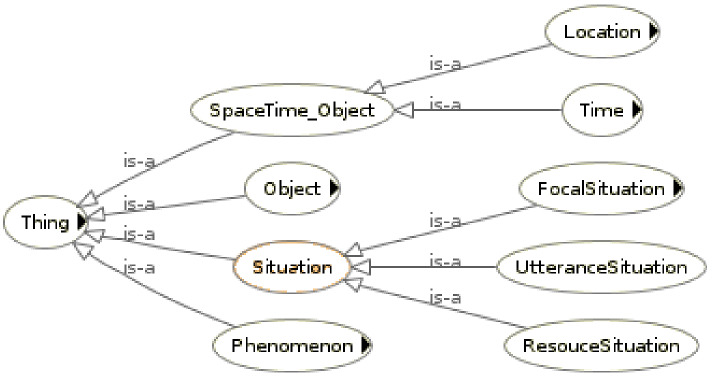


Figure 2.3: The generic part of STOR.

The generic ontology consists of concepts which are valid for all the domains. All of these concepts have subclasses which are specific for a domain and part of the domain ontology. The generic ontology consists of the concept *Situation*, *Object*, *Phenomenon*, *SpaceTime_Object* (with subclasses *Location* and *Time* which represent the same concepts as in Situation Theory). The concept *Individual* and its subclass *RealIndividual* is omitted from STO and in STOR represented as instances of a concept which is more in line with representation as designed with OWL.

We distinguish in STOR three types of *Situation*: *FocalSituation*, *ResouceSituation* and *UtteranceSituation*, as is done in Situation Theory and STO. The *UtteranceSituation* is interpreted differently in STOR than in STO and in conformance with Situation Theory. In STO an *UtteranceSituation* is used to denote the situation of the user of a situation awareness system which incorporates STO. An utterance in such a system is the utterance of a user who wants to generate information and, as such, queries the system. Such a user may be interested in a specific location and retrieves all the infons related to that location from the system. In STOR an *UtteranceSituation* is the *Situation* in which information is provided. This difference gives us the opportunity to annotate the information with additional data. When, for example, someone reports about a *FocalSituation* from a long distance, such information should be used differently than a report from someone who is nearby the same *FocalSituation*. The difference can be found by determining the *UtteranceSituation* which is supported by infons showing the location. Because we are working in the domain of crisis management further information can also be of interest such as the physical condition of the speaker and his relation with the *FocalSituation*. The speaker is known in the system as a *Person*, i.e. a subtype of *Object*, within a specific *UtteranceSituation*.

The difference between *Object* and *Phenomenon* is that the subclasses of *Object* are simple concepts such as *Ship*, *Building* or *RoadVehicle* while the

Figure 2.4: An example of the domain part of STOR.

subclasses of *Phenomenon* are complex concepts such as *Sight* or *Weathertype*. These complex concepts are often more difficult to determine than subclasses of *Object* because they are more vague or even subjective. Subclasses of the concepts *Object* and *Phenomenon* are part of the domain ontology. *Location* and *Time* are part of the generic ontology because in every domain specifications of locations and categorization of time occur. The generic part of STOR is shown in Fig. 2.3.

Concepts which are part of the domain ontology are specific for a particular domain. A knowledge representation of crisis management consists of *CrisisSituation* and concepts representing objects that are part of the crisis. In Fig 2.4 several examples of *CrisisSituation* are shown such as *CarAccident* and *Fire*. A representation of a specific domain can be more fine grained when appropriate. Subclasses of *Object* are a representation of individuals being part of a *Situation*. Examples of these representations are *Streetobject*, *Building* or *Person*. Instances of these concepts refer to real objects or people which somehow have a relation with the *Situation* we want to determine.

In a domain several categorizations for *Location* and *Time* can be appro-

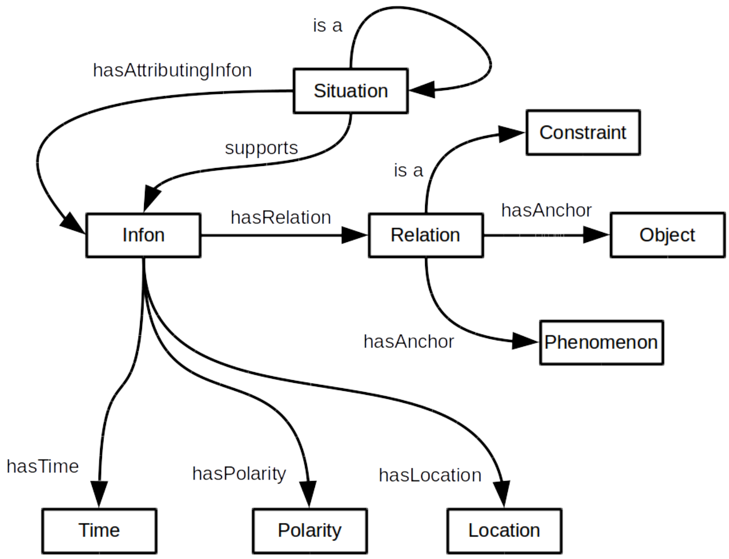


Figure 2.5: Properties of concepts in STOR.

priate. For example, when we want the location of a mobile phone user the GPS-coordinates are appropriate, but when we ask someone to tell us where he is, concepts such as *near* and subclasses of *Landmark* should be part of the ontology. Subclasses of *Time* are also domain-specific and depending on what kind of categorization of time is appropriate.

In STOR a *Situation* has several properties. A *Situation* is first of all a *Situation* of some sort, i.e. *FocalSituation*, *ResourceSituation* or *UtteranceSituation*. Two properties specify the *Situation*: *supports* and *hasAttributingInfon* which respectively define the *Situation* and provide further but non-essential information. Furthermore, a *Situation* can be part of a constraint as an antecedent or consequent. An *Infon* has four properties: *hasTime*, *hasLocation*, *hasPolarity* and *hasRelation*. *Relation* has properties *hasAnchor* which determines what kind of *Objects* and *Phenomena* are involved. The relations are shown in Fig. 2.5.

2.4 Discussion

The basic terminology of Situation Theory and OWL differ somewhat. What Situation Theory calls an object is referred to in OWL as an entity. So in STOR the term object is restricted to physical and abstract objects, distinct from situation and relations. Individuals in Situation Theory are mapped to objects in STOR. Real individuals, i.e. unique entities in the world, are represented as instances of OWL objects.

To map Situation Theory to OWL we represented the types and key entities such as infon and situation as classes. *Infon* and *Situation* have an internal structure analogous to the structure suggested by Situation Theory. Relations such as *supports* and *hasAttributingInfon* are represented as properties of classes. Parameters in Situation Theory are represented in OWL as value restrictions of properties.

Both object type and situation type are represented in OWL as classes, but situation type classes have internal structure, i.e. supported infons, and object type classes do not necessarily have internal structure. So object types can be represented in an OWL *is-a* hierarchy without further definitions. For example, object types such as *ship* and *building* are represented as simple classes in the object hierarchy. Situation types are represented in OWL as subclasses of *situation* and have one or more *supports* properties with (parametric) infons as value. These infons classify a given situation as being of some situation type. For example, a *CarAccident* has two supported infons *InjuredInfon* and *MultipleVehicleInfon*. Both of these infons are parametric due to an abstraction over the polarity which makes *CarAccident* an abstraction of four different situations.

Constraints are represented in OWL as instances of the entity *Constraint*, which has two properties: *antecedent* and *consequent*. The values of these properties are situation types. For example, the situation *CarAccident* has two attributing infons: *WeathertypeInfon* and *StreetConditionInfon*. When the weather type is *Rain* the consequent is a situation where the street condition is *Wet*.

We represent three of the four types of constraints defined in Situation Theory in STOR. The linguistic constraint is not explicitly mentioned. The linguistic constraint defines the relation between terms used to denote things and these things in de real world, e.g. when someone says ‘building’ he is referring to something in the real world. This reference is defined by a linguistic constraint. We do not represent this constraint in STOR because it is represented by instances of *IND* and their association with objects in the real world.

The dual association of each entity with both a type, i.e. a subclass of *TYP*, and a superclass seems somewhat redundant. Indeed the type can be inferred from the position of an entity in the *is-a* hierarchy. However, for the classifier it can be informative to know the type of an entity even if its position in an *is-a* hierarchy is not yet known.

We made a clear distinction between representational, generic and domain ontology which was lacking in STO. This distinction results in an ontology which is more robust and scalable than STO.

2.5 Conclusion

In this chapter we introduced Situation Theory as a formal system to capture situations. We use situations to determine and communicate about what is going on in the real world. Situation Theory offers various forms of abstraction which can be used to generate questions as will be elaborated in the next chapter. With Situation Theory Ontology Revised the knowledge about what is going on in the real world can be captured and reasoned about, in a widely used format, i.e. OWL.

Chapter 3

Situation Awareness Question Generator

The theoretical framework we created in the previous chapter formalizes situations, but it does not specify the order in which questions should be asked. An implementation of the framework and question strategies are developed for a system called 'Situation Awareness Question Generator'. This application makes use of an ontology which represents several situations and is used in experiments, as presented in the subsequent chapters. Parts of this chapter were published as two conference papers (Teitsma et al., 2011a,b).

3.1 Introduction

In this chapter we develop several strategies to retrieve information from people about a crisis situation as fast as possible. By asking automatically generated questions we retrieve information from a large group of people, which is known as 'crowd sourcing' (Howe, 2006). The communication with the users of our system is done via the mobile device most people carry with them most of the day. We present the Situation Awareness Question Generator (SAQG) which uses the information retrieval strategies to determine the certain aspects of a crisis situation.

The use of ontologies for automatically generating questions about a crisis situation which are asked to a large number of people requires that the number of questions is minimal, the questions as informative as possible and easy to answer. In this chapter we try to enhance this by formulating strategies for asking the right questions in the right order. In subsequent chapters we elaborate on the trustworthiness of the information thus gathered.

The formalization of situations using Situation Theory and Situation Theory Ontology as done in Chapter 2 gives rise to an enormous amount of questions to be asked when we want to determine a situation: every possible abstraction is a potential question. Should we ask all those questions in a random fashion without any plan, determination of the situation we are interested in would take a long time. In case of an emergency this would take too long. To shorten the time for determination of the situation a selection and ordering of the most informative questions has to be made. For the selection and ordering we use the informational and truth value of possible answers, i.e. the amount of information of an answer and whether these answers can be true or not, to compute the best question at a given moment. The best series of questions is the sequence of questions which renders the most information with the least number of questions about a particular situation.

The informational value of a question is computed using a theory of information which offers concepts to distinguish between informational entities. The degree of abstraction and degree of accuracy give a value to the informational entities representing the amount of information of such an entity. The most efficient question is that question which, when answered, reduces the discrepancy with the truthful description of the situation the most. An answer to such a question should always generate information whether it is a positive or negative answer when asking a polar question or a specification when asking a multiple choice question. By asking for infons supported by a situation we determine situations using several strategies based on properties of the ontologies describing the domain.

In Section 3.2 an elaboration of Semantic Information Theory shows how the informational and truth value of (compound) infons can be computed. In Section 3.3 several strategies to find the question which gathers the most information at that moment are presented. With the theoretical framework as presented in the previous chapter and supplemented by a Semantic Information Theory a communication system is developed. This system is discussed in Section 3.4.

3.2 Theory of Strongly Semantic Information

The amount of information of a statement is computed by referring to the probability of specific information, according to the Semantic Theory of Information (Carnap and Bar-Hillel, 1952) and the Mathematical Communication Theory (Shannon, 1948). The probability of an informational entity such as a statement or a compound statement, being true depends on what is not being stated, according to the scholastic dictum *omnis determinatio est negatio (every determination is a negation)* (Bremer and Cohnitz, 2004) also known as the Inverse Relationship Principle (D'Alfonso, 2012). This principle states that the amount of information of a statement is inversely related to the probability of this statement, e.g. when it is highly unlikely some statement is true, it is highly informative to know that it is true.

The Inverse Relationship Principle implies that an analytic statement, i.e. a statement which is true by the meaning of its components, has the least possible informational value because it is always the case that such a statement is true. When a possible situation is always true its probability is 1 and the informational value becomes 0 ($log_n(1) = 0$). And reversely, when a possible situation has a probability of 0 as in a contradiction, then the informational value becomes infinite ($log_n(0) = \infty$).

Floridi criticizes the idea that a contradiction gives the greatest possible informativeness to a statement (Floridi, 2004). In his Theory of Strongly Semantic Information he solves this problem by stating a correspondence between how things actually are and the informativeness of an infon, i.e. the informational value of an infon is computed by referring to the discrepancy between the

	σ_1	σ_2	σ_3	σ_4
S_1	0	0	0	0
S_2	0	0	0	1
S_3	0	0	1	0
S_4	0	0	1	1
S_5	0	1	0	0
S_6	0	1	0	1
S_7	0	1	1	0
S_8	0	1	1	1
S_9	1	0	0	0
S_{10}	1	0	0	1
S_{11}	1	0	1	0
S_{12}	1	0	1	1
S_{13}	1	1	0	0
S_{14}	1	1	0	1
S_{15}	1	1	1	0
S_{16}	1	1	1	1

Table 3.1: Table with possible situations when having four infons.

degree of truthfulness or specificity of the infon and the complete truthful and specific description of the situation. Floridi's Theory of Strongly Semantic Information is based on truth values and thereby precludes the Bar-Hillel-Carnap semantic paradox in which a contradiction has more information content than a true statement (Floridi, 2011). To compute the informativeness of statements Floridi takes two factors into account:

- the truth value of the statement and
- the degree of discrepancy between the statement and the actual situation.

With these two factors Floridi distinguishes between truth and falsehood in various degrees with respect to a particular situation.

A situation is often described using more than one infon and as such creating a sentence. A sentence is a description of a situation constructed as a number of infons which are connected by conjunctions or disjunctions. An example of a situation is the flooding of a region which is described, in this abstract situation, by a sentence consisting of four infons:

$$\begin{aligned}
&\sigma_1 = << flooded, streets, Oude\ Tonge, 01021953, \dot{p} >> \\
&\sigma_2 = << rising, water, Oude\ Tonge, 01021953, \dot{p} >> \\
&\sigma_3 = << storm, NorthbyNorthwest, Oude\ Tonge, 01021953, \dot{p} >> \\
&\sigma_4 = << evacuated, people, Oude\ Tonge, 01021953, \dot{p} >>
\end{aligned} \tag{3.1}$$

where $\dot{p}$ is a parameter of type POS (i.e. it indicates the polarity: 0 or 1 (see chapter 2). With four infons we can distinguish 16 situations of which only one denotes the actual situation. All the other descriptions, as shown in Table 3.1, are false descriptions of the actual situation and called 'inaccurate'. Another way to describe the situation is to make sentences which are true but are not the most specific description of the situation. These descriptions refer to more than just the actual situation and are called 'vacuous'.

3.2.1 Inaccurate Descriptions

With respect to the truth value of a sentence we can say that a sentence is false when one of the infons used to describe the situation does not correspond to what is going on in the world. When, for example, it is stated that the streets in Oude Tonge are *not* flooded on February 1st 1953 while in reality they were flooded with water a meter high, this infon is false and as a consequence the description of which it is part is also false. When a sentence is false, relative to the actual situation, this sentence is called inaccurate to some degree. A sentence can be more or less inaccurate when parts of it are an accurate description of the actual situation and other parts are not.

The degree of inaccuracy, referred to as υ_i, with respect to the situation under observation S is calculated as the ratio between the number of erroneous infons e and the number of infons supporting the possible situation $n(\sigma)$ (the total number of infons describing the possible situation)[1]:

$$\upsilon_i(S) = -\frac{e(S)}{n(\sigma)} \tag{3.2}$$

In the context of a situation supported by four infons, as shown in Table 3.1 this will give 4 classes of inaccuracy as shown in table 3.2.

Nr. of erroneous infons	Degrees of inaccuracy
1	-1/4
2	-1/2
3	-3/4
4	-1

Table 3.2: Classes of inaccuracy.

For example, when in a description of a situation consisting of four infons, one infon is not accurate then this description contains more information than a description of the same situation containing two inaccurate infons.

3.2.2 Vacuous Descriptions

An infon which corresponds with the actual situation can be true even when it is not the most specific description of this situation, i.e. an infon can refer to more than only the actual situation and is thus an abstract infon. When, for example, is stated that the people of Oude Tonge are evacuated and the streets are flooded, but it is not known whether the water is rising and whether the storm is coming from North by Northwest, this sentence is true but refers to three possible situations, one of which is the actual situation. When a sentence is true but too abstract, it is called vacuous to some degree. A sentence can be vacuous to a certain degree when it refers to an abstract situation which encompasses the actual situation.

The most vacuous sentence is the sentence which refers to all possible situations, i.e. the tautology. This sentence does not inform at all about the actual situation. When one fact is known about the actual situation, i.e. it is known

[1]Floridi multiplies e with -1 with the end to combine it with the formula of vacuity, i.e. equation 3.3.

that one of the infons in the sentence is correctly describing a part of the actual situation, a sentence is less vacuous. Such a sentence is more informative than the tautology. To calculate the vacuity Floridi uses the number of situations referred to by the sentence and the total number of possible situations given the number of infons and their number of truth values:

$$\upsilon_v(S) = \frac{n}{n(S)} \tag{3.3}$$

where n is the cardinality of the set of all the situations which are referred to by the sentence, necessarily including the actual situation and $n(S)$ the total number of possible situations. The degrees of vacuity as formulated by Floridi are computed under the assumption that the infons are independent of each other. The degree of vacuity can also be computed by the formula

$$\upsilon_v = \frac{\frac{n(S)}{2^c} - 1}{n(S)} \tag{3.4}$$

where c stands for the number of conjunctions and $n(S)$ for the total number of possible situations.

Nr. of disjunctions	Nr. of situations	Degrees of vacuity
3	15	15/16
2	7	7/16
1	3	3/16
0	1	1/16

Table 3.3: Classes of vacuity.

With semantic weakening a series from a minimum vacuity to maximal vacuity is created in a systematic way. Semantic weakening is done by connecting the infons which constitute the situation by gradually more disjunctions instead of conjunctions. When we make the statement $\sigma_1 \wedge \sigma_2 \wedge (\sigma_3 \vee \sigma_4)$, it can be seen in Table 3.1 that the situations S_{14}, S_{15} and S_{16} support the statement. Situation S_{13} does not support the statement because σ_3 and σ_4 both are false and $\sigma_3 \vee \sigma_4$ does not result in a true statement. A statement with two disjunctions results in the infon $\sigma_1 \wedge (\sigma_2 \vee \sigma_3 \vee \sigma_4)$. This infon complies with even more situations: first of course S_{14}, S_{15} and S_{16}, and then also with S_{10}, S_{11}, S_{12} and S_{13}. When making the statement $\sigma_1 \vee \sigma_2 \vee \sigma_3 \vee \sigma_4$, all situations except S_1 are supported. This results in Table 3.3.

3.2.3 Informational Value

The degree of informativeness ι of a sentence σ can be calculated with the formula (Floridi, 2004):

$$\iota(\sigma) = 1 - \upsilon^2(\sigma) \tag{3.5}$$

In this formula υ is the discrepancy between the description under evaluation and the true description of the actual situation which can have a degree of inaccuracy (υ_i) or vacuity (υ_v). Both inaccurate and vacuous information is contingently true information, i.e. in this world it is possibly true. By taking both accurate and vacuous information into account, it becomes possible to compare inaccurate and vacuous information, e.g. it becomes possible to prefer an inaccurate description with a small discrepancy, to a vacuous description with a large discrepancy with an accurate and most specific description of a real situation.

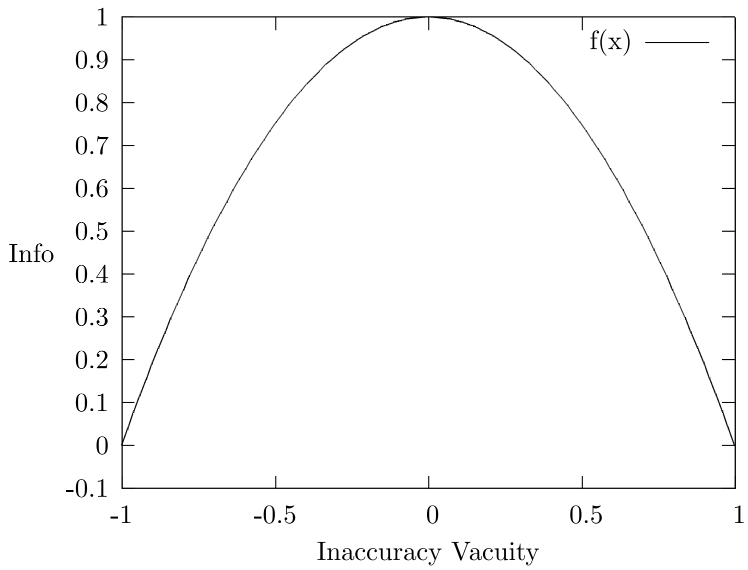


Figure 3.1: The informativeness of a sentence, i.e. compound infon (σ). $f(x) = \iota(\sigma)$. From: Floridi (2011).

The graph shown in Figure 3.1 represents the informativeness of sentences. In this figure the x-axis represents the discrepancy of σ from the true description of the actual situation w which has discrepancy 0. The y-axis is the result of formula 3.5. Note that the negative discrepancy is the field of inaccuracy and the positive discrepancy is the field of vacuity. When a sentence consists of a contradiction it can never be true, has the greatest inaccuracy and thus the least informativeness. Maximal vacuity for a statement corresponds to a tautology in a specific domain because it is always true. Decreasing the number of situations which are compatible with the true situation increases the quantity of informativeness. A statement has a minimum vacuity when it refers to the minimum number of situations and thus has the greatest informativeness.

3.3 Strategies for Question Generation

In general, situations are described with multiple infons and several questions have to be asked to determine a situation. When asking a series of questions one has to select which questions to ask first, second, etc. Several strategies can be formulated to automatically generate a series of questions.

The questions we ask, are specified by the description of the situation we want to determine and what is considered important to know about this situation. Also the level of detail does influence this sequence. For example, sometimes it is sufficient to know whether someone is injured or not, without knowing the specific injury.

We developed several strategies because of differences in the description of situations as represented in ontologies in the domain. Situations are described using multiple infons which may have relations with each other or not. When there are relations, the properties of these relations influence the choice of the specific strategy we can use.

3.3.1 The First Strategy

The first and most simple strategy is a strategy which asks after each possible situation. This strategy is used when no assumption about the description of the possible situations is made. Such is the case when the situations are described with different infons, i.e. the infons which are supported by each situation do not relate to each other in any way. When described in a formal manner, all situations are different from each other. For example, a situation can be defined using four infons: ‘the streets are flooded’, ‘the water is rising’, ‘the storm is North by Northwest’ and ‘people are evacuated’, while another situation is defined using three infons: ‘it is raining’, ‘there is a storm’, ‘people are walking on the street’. With this strategy the number of questions one should ask potentially equals the number of possible situations minus one. The questions asked are of the form: ‘Are the streets flooded, is the water rising, is the storm North by Northwest and are people evacuated?’ *and* ‘Is it raining, is there a storm and are people walking on the street?’. Whether the questions asked, are approaching the description which corresponds to the actual situation is not known and when it is, this is pure coincidence and not the consequence of a deliberate strategy other than to use the order of some list. The inaccuracy of each description is not further investigated, i.e. whether the inaccuracy is high or low is not asked for and not taken into account when choosing the next question. With this strategy a representation of information gain per question based on information value is not possible. The only thing one can say is that each time a question is asked, it will take less time to get the right description of the actual situation.

Algorithm 1 First strategy.

```
retrieve the subordinate concepts of FocalSituation from the ontology and
make them element of a set A
while set A is not empty do
    get the infons which are supported by the subordinate concept of FocalSi-
    tuation (an element of A)
    concatenate the infons with conjunctions and assign it to question
    ask whether the answer to the question is true or false
    if the answer is true then
        mark the situation as actual
        end the while loop
    end if
    if the answer is false then
        remove the situation which is used from set A
    end if
end while
```

This strategy is represented in Algorithm 1. Firstly, all the subordinate concepts of FocalSituation are retrieved and made element of a set. Then the infons supported by that situation are retrieved and shown as representing the situation. These infons are concatenated and subsequently shown to the user of the application with the question 'Is this true?'. For example,

```
Is it true that:
        the streets are flooded and
        the water is not rising and
        the storm is North by Northwest and
        people are not evacuated.
```

When the description of a situation which corresponds to the actual situation is shown and the user chooses 'yes' the while-loop is ended. This information is send to the server and the application is stopped. We did not use this strategy in our application because it is too laborious and other strategies are faster in determining the situation with the ontologies we use. The efficiency is not high: the time needed to find the description of the actual situation is proportional to the number of possible situations.

3.3.2 The Second Strategy

Except for the first strategy, we use Floridi's Theory of Strongly Semantic Information to find the best strategy given the characteristics of the description of the situations. We always start with a sentence which is maximally vacuous. Dependent on the description of the situation and the characteristics of the ontology we look for the question which brings us, whatever answer is chosen, closer to the description of the actual situation.

The second strategy assumes that all situations are described with the same set of independent parametric infons. An infon is independent from another infon when there is no relation between these infons, i.e. the accuracy of one infon has no influence on the accuracy of another infon. For example, the color of an object is independent of the shape of an object. The two infons describing the color and shape are then also independent from each other. Here we ask for each specific infon which is supported by the situation. To determine a situation we always ask for a parametric infon. The parameter of this infon is restricted by conditions which denote the possible informational entities which satisfy the conditions. The parameter is always of the basic type IND which is the type denoting individual objects (see Chapter 2). In practice this is done by presenting multiple choice questions to the user of our application. An example is the visual description of a person. A person can have one of several hair colours and this person can wear several types of clothing. Such infons are independent from each other and to describe a person one necessarily has to ask for each infon. The number of questions is equal to the number of infons which are supported by the situation. It is clear that this strategy is superior to the first strategy with respect to the number of questions: when all the possible situations should be asked about, the number of questions grows with the number of referents per infon describing the situation. At first the description is maximally vacuous and after each answer an infon is specified and the description of the situation becomes less vacuous.

Using this strategy, the information gain per question will depend on the number of possible answers, i.e. the number of referents of the parameter which are mutual exclusive. When, for example, there are three parametric infons, $\sigma_1, \sigma_2, \sigma_3$, describing the situation and the parameter used in σ_1 has eight referents, the parameter used in σ_2 has five referents and the parameter used in σ_3 has four referents, this results in 160 possible situations. In a formula this is represented as:

$$(\sigma_{1a} \vee \sigma_{1b} \cdots \vee \sigma_{1h}) \wedge (\sigma_{2a} \vee \sigma_{2b} \cdots \vee \sigma_{2e}) \wedge (\sigma_{3a} \vee \sigma_{3b} \cdots \vee \sigma_{3d}) \tag{3.6}$$

The question which generates the greatest information gain is the question with the largest number of possible answers, i.e. the largest set of referents of the parameter. That is, in this example, the question asking for σ_1. When the question based on σ_1 is answered the reduction of the discrepancy with the most precise description of the actual situation will be 87,5% $(1 - \frac{20}{160})$. When each time a question will be asked the question with the greatest information gain is chosen, then the information gain is represented more or less as in Figure 3.1. The number of questions is equivalent to the number of parametric infons.

Algorithm 2 Second strategy.

get subordinate concepts of FocalSituation
get the infons which are supported by the FocalSituations and make them element of a set A
while set A is not empty **do**
  get the attribute of the infon and its subordinate concepts
  assign the attribute to question and the subordinate concepts to possible answers
  ask question and present possible answers
  get the answer and mark the infon as actual
  remove infon which is used from set A
end while

The second strategy is used when a situation consists of infons which have multiple possible answers (see Algorithm 2). The possible answers are elements of the set of subordinate concepts of the attribute which represents the parameter in the parametric infon. An example is the infon representing the weather, which is not yet known, at Oude Tonge on the 1st of February 1953:

$$\sigma = <<\dot{w}, Oude\ Tonge, 01021953, 1>> \tag{3.7}$$

where

$$<<\dot{w} = \text{Weathertype}_1 \upharpoonright << actualWeather, Oude\ Tonge, 01021953, 1>> \tag{3.8}$$

and

$$<< \text{Weathertype}_1 \in \{\text{Dry, Rain, Hail, Mist, Snow, Weathertype unknown}\} >> \tag{3.9}$$

which states that the weather at Oude Tonge on the 1st of February 1953 is of some sort. This weather is restricted by an abstraction called 'Weathertype' which has to satisfy the condition of being the actual weather at Oude Tonge on the 1st of February 1953 (see Chapter 2) and is restricted to a number of elements of a set. In ontology 'Weathertype' is a concept which has as subordinate concepts 'Dry', 'Rain', 'Hail', 'Mist', 'Snow', 'Weathertype unknown'.

These concepts are presented as multiple choices for an answer to the question after the weather type as part of the description of the actual situation. For example:

```
What is the weathertype?
       Dry
       Rain
       Hail
       Mist
       Snow
       Weathertype unknown
```

When one of the suggested answers is chosen, the parameter is replaced by the specification and the infon marked as actual. The efficiency of this algorithm is proportional to the number of parametric infons supported by the situation.

3.3.3 The Third Strategy

A third strategy is to ask for the infons which are supported by a situation, just as the second strategy, but now the number of possible answers is restricted to two possible answers: 'yes' or 'no'. The constellation necessary for this strategy consists of situations which are described with infons of which the polarity is unknown. An example is the infon

$$\sigma = \ll injured, people, Oude\ Tonge, 01021953, \dot{p} \gg \tag{3.10}$$

In a constellation as shown in Table 3.1 all four infons have a parameter of the type POL which restricts the reference to the set $\{true, false\}$. All possible situations are described with the same set of four infons. The possible answers

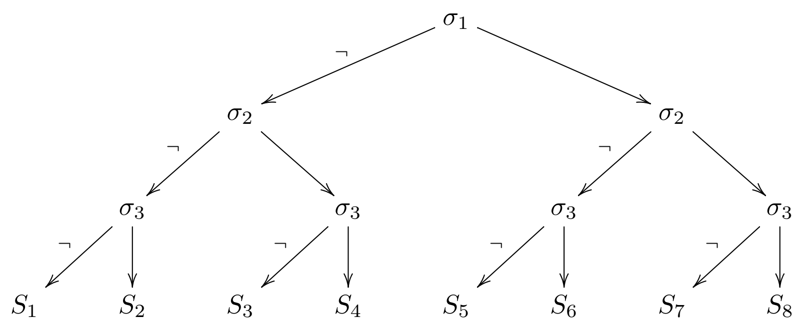


Figure 3.2: Decision graph.

for a question such as 'Are people injured?' are 'yes' and 'no' which both have the same informational value. In this constellation it does not matter which question is asked first. The number of possible situations is 2^n where n is the number of infons supported by the situation. Because we ask for the infons directly instead of the possible situations this strategy is more efficient than the first strategy. This strategy finds a question by binary search and is also known as 'divide and conquer' (Sedgewick, 2011). In a situation with three infons this would create a decision graph as shown in Figure 3.2.

Algorithm 3 Third strategy.

get a subordinate concept of FocalSituation
get the infons which are supported by this situation and make them element of a set A
while set A is not empty **do**
  ask question and present the answer 'yes' or 'no'
  get the answer and mark the infon as actual or not
  remove infon which is used from set A
end while

The representation of the information gain using this strategy is exactly as in Figure 3.1. The reason for this analogy is of course that this strategy is the reversal of semantic weakening as described in Section 3.2.2. While semantic weakening starts with the right description of the actual situation and gradually diminishes the number of conjunctions, we are, using this strategy, starting with the most vacuous description and then diminishing the number of disjunctions. The number of questions to ask and to determine the situation is equivalent to the number of infons used to describe the situation.

A representation of the third algorithm is shown in Algorithm 3. An example is the situation in which several cars are involved and people are injured during an accident. In OWL this situation is defined as follows:

```
Class: CarAccident

    EquivalentTo:
        supports min 1 InjuredInfon,
        supports min 1 AccidentWithCarInfon

    SubClassOf:
        STOR:CrisisSituation
```

The situation *CarAccident* supports two infons, i.e. *InjuredInfon* and *AccidentWithCarInfon* which both are a subclass of *ParametricInfon.* Because the polarity of this infon is not determined, the resulting question is a polar question, i.e. the answer consists of 'yes' or 'no'. For example:

```
Is it true that people are injured?
```

The answer tells which description of the actual situation is correct. Of course, in this case, one should ask also a question whether one or more vehicles are involved to determine the right and most precise description.

3.3.4 The Fourth Strategy

The fourth strategy we call 'semantic strengthening' and is analogue to the third strategy but is further restricted by the assumption that in practice certain situations are prohibited. This strategy is used when some theoretically possible but in practice impossible situations are involved because the infons are *not independent* from each other. An example such as in Table 3.1 with a situation supporting four infons, is an actual situation in which streets are flooded, people are sailing by boat, it is cold and raining:

$$
\begin{aligned}
&\sigma_1 = \ll \mathit{flooded}, \mathit{streets}, \mathit{Oude\ Tonge}, 01021953, \dot{p} \gg \\
&\sigma_2 = \ll \mathit{sailing}, \mathit{boat}, \mathit{Oude\ Tonge}, 01021953, \dot{p} \gg \\
&\sigma_3 = \ll \mathit{cold}, \mathit{Oude\ Tonge}, 01021953, \dot{p} \gg \\
&\sigma_4 = \ll \mathit{raining}, \mathit{Oude\ Tonge}, 01021953, \dot{p} \gg
\end{aligned}
\tag{3.11}
$$

The theoretical information space becomes smaller when confronted with the rules of a specific domain. In the context of crisis management categorization of flooding restricts the theoretical information space by declaring some situations impossible. Suppose it is impossible to sail by boat when the streets are not flooded. This makes the situations S_5, S_6, S_7, S_8 impossible. The first strategy is still applicable, but when using the third strategy the problem is that this strategy may result in an impossible outcome: when the answer to the question after σ_2 is affirmative and the answer to the question for σ_1 is negative the resulting compound infon would be 'a boat is sailing in a street which is not flooded'. The strategy has to prohibit such impossible descriptions while determining the actual situation. This strategy takes constraints (see Chapter 2) into account. Such a constraint, in this example, would be 'if streets are flooded then a boat can sail in Oude Tonge'. Constraints represent dependencies in the description of the situation, i.e. the acknowledgement of a specific infon is only allowed when another infon is also acknowledged.

These dependencies between infons are formalized in the ontology we use for our application. Such rules are of the form of an implication between an antecedent and consequent. With these rules some questions become redundant and do not have to be asked. The order of questions again is defined by the information value of the questions. The distribution of the information value is different from a constellation when there are no impossible situations. To avoid an impossible situation we first ask the question for which the answer implicates an answer for another question. We use an algorithm which generates questions for which the answers result in a vacuous description until the right description of the actual situation is given, i.e. it never generates questions for which the answer is an inaccurate infon. For this we make a table with only possible situations and use this table to compute the information value of each question. The number of questions is the same as for the third strategy minus the number of questions which are not asked because by implication the answer is already known.

The information gain for each question differs depending on whether an answer implicates other infons. When an answer implicates another infon the information gain is larger than when an answer does not implicate another infon:

$$
\left.\begin{array}{r} \neg\sigma_1 \rightarrow \neg\sigma_2 \\ \neg\sigma_1 \end{array}\right\} \neg\sigma_2
\tag{3.12}
$$

Before gathering an answer one does not know what the information gain exactly will be. The strategy is to ask the question with potentially the greatest information gain. But it is still possible to get an answer which does not imply another infon. Therefore, it is not possible to tell in advance how the growth of information will be realized. However when an implication is activated, the information gain is larger than when an implication is *not* activated. It is clear that in the case of an activated implication a line representing the information gain will be discontinuously progressive instead of continuously progressive as in Figure 3.1.

The fourth strategy is shown in Algorithm 4. The first step in this strategy is to identify and remove the impossible situations. This is done by retrieving the infons used by the subordinate concepts of FocalSituation and making a table representing all theoretical possible situations (cf. Table 3.1). Then the rules are applied and the impossible situations identified. These impossible situations are eliminated from the table as possible outcome of the sequence of questions and answers. Then we ask for the infon which is most common among the possible situations.

Algorithm 4 Fourth strategy.

```
get subordinate concepts of FocalSituation
get infons used by these concepts
create all the theoretical possible situations
get the rules for this specific domain
eliminate the impossible situations
while set A is not empty do
  get the most common infon
  ask question and present the possibility to answer 'yes' or 'no'
  get the answer and mark the infon as actual or not
  remove infon which is used from set A
end while
```

An example of a question using this strategy is similar to the example presented with the third algorithm. The difference is not found in the questions themselves but in the relation between questions. Sometimes a question does not have to be asked because by implication it is already answered.

3.3.5 The Fifth Strategy

The fifth strategy searches for the most detailed description available in the ontology. In this constellation an ontology represents knowledge of a specific domain in a hierarchy of concepts which are related to each other by 'is-a relations'. Superordinate concepts have subordinate concepts which specify characteristics of the superordinate concepts. All the questions are multiple-choice questions, i.e. the question asks for a further specification and the possible answers are presented to choose from. An example is the question 'What is on fire?' with possible answers: 'Building', 'Industry', 'Nature', 'Road vehicle', 'Aircraft', 'Ship', 'Rail vehicle' and 'Street object'. When, for example, 'Building' is chosen a further question is generated: 'What kind of building?' with as possible answers the subordinate concepts of 'Building'. This sequence of questions stops when the concept has no subordinate concepts in the used ontology and the most specific description of the object under observation is found or when the user can not make a further specification.

This strategy is used when we have an ontology which is composed of *is-a* relations. Each superordinate concept has subordinate concepts which are a specification of the superordinate concept. For example, a building can be an apartment building, a bungalow, an office building and many more types of building. Using this strategy we start at the most encompassing concept and keep asking questions until the most detailed description is given. The answers to these questions always are more or less vacuous infons, except for the final answer.

The questions asked, are represented as follows:

$$\sigma_{1a} \vee \sigma_{1b} \cdots \vee \sigma_{1\nu} \tag{3.13}$$

where ν represents the last referent of the parameter in the infon. The user chooses a specific infon as representing best the actual situation and is presented with another series of infons. The process continues until the most specific concept from the ontology is presented. The specific informational gain with each question depends on the specific number of subordinate concepts of each concept. In a more or less balanced ontology the relative informational gain at the start of the series of questions is the greatest because then the largest number of otherwise possible situations is disregarded. This relative gain diminishes with each question. A graphical representation of this strategy will be close to or similar as shown in Figure 3.1 but again the precise information gain can vary. The number of questions is determined by the number of parametric infons used to describe a situation and the number of *is-a* relations from the top of the hierarchy of concepts to the bottom, i.e. the most specific concept. The number of questions coincides with the path length of the particular subtree.

A representation of this strategy is shown in Algorithm 5. The algorithm starts at the top of a hierarchy of concepts with the most superordinate concept and its subordinate concepts. The subordinate concepts are presented in a multiple-choice list as possible answers to the question for a further specification of the superordinate concept. When the user chooses a specific concept from the list of presented subordinate concepts, then this concept becomes the concept which we want to specify. This procedure is repeated until the most specific concept is chosen.

Algorithm 5 Fifth strategy.

```
get the most superordinate concept and assign this to C
while concept C has subclasses do
  get the subordinate concepts of concept C and assign these to set S
  ask question and present all elements of S as possible answers
  get answer and set answer as concept C
end while
substitute the last answer for the parameter and mark the infon as actual
```

An example of a question generated by this strategy is similar to a question generated with the second algorithm. The difference with the second algorithm is that the question which comes next is related to the first question. The questions more and more try to specify the concept which was chosen earlier. An example is:

```
What kind of building is on fire?
        Building for entertainment
        Building for education
        Bunker
        Station
        Monument
        Residential building
```

And when for example 'Residential building' is chosen then the following question is asked:

```
What kind of residential building is on fire?
        Multiple floor building
        Building for needy
        Holiday accomodation
        Home
        Premises
```

When the ontology for one of the concepts shown does have subordinate concepts, these subordinate concepts are shown when that concept was chosen until a chosen concept does not have any subordinate concepts.

3.3.6 Summary

The definition of the situation we want to determine influences the strategy we use. The definition of a situation is the result of how the domain of interest is represented. The presented strategies vary according to whether the infons used have parameters with references from the type POL or IND and whether there are rules to follow. When the domain consists of unrelated facts with a great variety, the second strategy is of use. When rules in the domain are important and the infons are mutually dependent, one should use the fourth strategy. When situations are described with a small set of infons which are independent from each other, the third strategy is preferred. And when one is looking for a specific description and an ontology describing the taxonomic relations between concepts is available, the fifth strategy should be used. In Table 3.4 a scheme is presented which shows the preferred strategy when taking the type of reference and the presence of rules into consideration.

	Type of reference	
	Individuals	Polarity
No rules	2	3
Rules	5	4

Table 3.4: Scheme of strategies.

It is also possible to use these strategies in combination, when a situation is described using more than one of the mentioned characteristics. For example, when formalizing a *Car Accident* some parametric infons have parameters referring to 'yes' or 'no' and others to a set of concepts which generate respectively yes-no and multiple-choice questions.

3.4 Architecture of the System

Our system consists of two sides: a server and a client. The server sends ontologies to a client which uses the ontologies to produce questions. The answers to these questions are preserved until a situation is determined. Then the answers are send to the central server and possibly used for further computation. In the model we propose such a server has a repository of several domain ontologies.

The scenario we envision starts with the call of a citizen who is in distress about some situation. Nowadays the citizen calls the emergency service number 112 (or 911 in North America) and when the nature of the emergency is determined, he will be redirected to the police, fire or medical department is done. In our scenario the citizen has an application on his or her mobile device and is redirected to the right ontology, i.e. police, fire or medical, and answers questions generated from the ontology. The ontologies are preserved on the mobile device but are regularly updated from a server. When a distressful situation occurs, the application is activated and the citizen answers the questions. When the user has answered the last question, the determined situation is sent to the server and handled accordingly, i.e. the appropriate emergency service is activated and receives the gathered information.

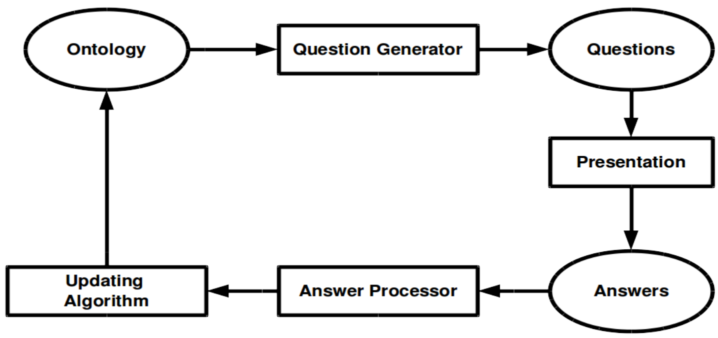


Figure 3.3: Situation Awareness Question generator (SAQG).

The application as used on the client-side is composed of several parts. Figure 3.3 shows the architecture of SAQG. An ontology is designed and used, with the help of semantic information theory, to generate the infon which is most interesting to ask after. The questions which accompanies that particular infon is asked. The question is answered by the user and the answer is then computed to update the ontology.

3.4.1 Question Generator

The reading, modeling and manipulation of the ontology we use to generate questions is done with Jena. Jena offers a comprehensive API to create functions handling the information stored in such an ontology (Carroll et al., 2004; Grobe, 2009). Before an algorithm to generate questions is used, first the ontology is retrieved and transformed into a Triple DataBase (TDB) for high performance. The TDB can be accessed in the same manner as the access to an ontology represented by OWL is done, using exactly the same queries, but the retrieval of data is faster.

As explained in Chapter 2, our ontology, i.e. STOR, consists of a representation, generic and domain-specific part. The representational and generic part of the ontology are seperated from the domain-specific part. The first and second part of the ontology are distributed with the application but the latter part is updated when opportune. In such a file, subclasses of *Situation*, *Object* and *Relation* are specified. This division between a general ontology and a more specific, domain-related, ontology makes it easy to extend the ontology and still keep exchange of information between the domain specific ontologies possible.

An example of the Graphical User Interface presented to the user of SAQG is shown in Figure 3.4. It is supposed here that the user has answered a previous question about what kind of emergency he wants to report with 'Something is on fire'. The follow-up question is to ask for what is on fire. The question is, in this case, generated by using the fifth strategy on a *is-a* hierarchy which represents a categorization of objects in the world.

To determine the situation, SAQG is looking for subclasses of *FocalSituation* and their supporting infons. Subordinate concepts of *FocalSituation* can become actual (see Section 3.3) and therefore are used to generate questions. In all the strategies infons which specify the situation are used to construct the questions in one way or another. Because each infon is an informational entity, it is easy to formulate a question asking precisely for this piece of information. Each infon then has a question as a label. This label can be read when an infon is chosen as the infon to ask for. When wanting to ask whether people are injured, reading of the label will give 'Are people injured?' An answer to this question is added as a fact to the ontology. When a complete situation is determined this information is send to the ontology residing on the server.

The algorithm that controls the sequence of events in SAQG is presented in Algorithm 6. When a situation is found to be actual, the questions corresponding to attributedInfons are asked and the answers become part of the actual situation.

3.5 Conclusion

In this chapter we have presented an application of the theoretical framework as elaborated in Chapter 2. With the Theory of Strongly Semantic Information we can make a difference between descriptions of a situation which are inaccurate, i.e. false descriptions and descriptions which are vacuous, i.e. true but too

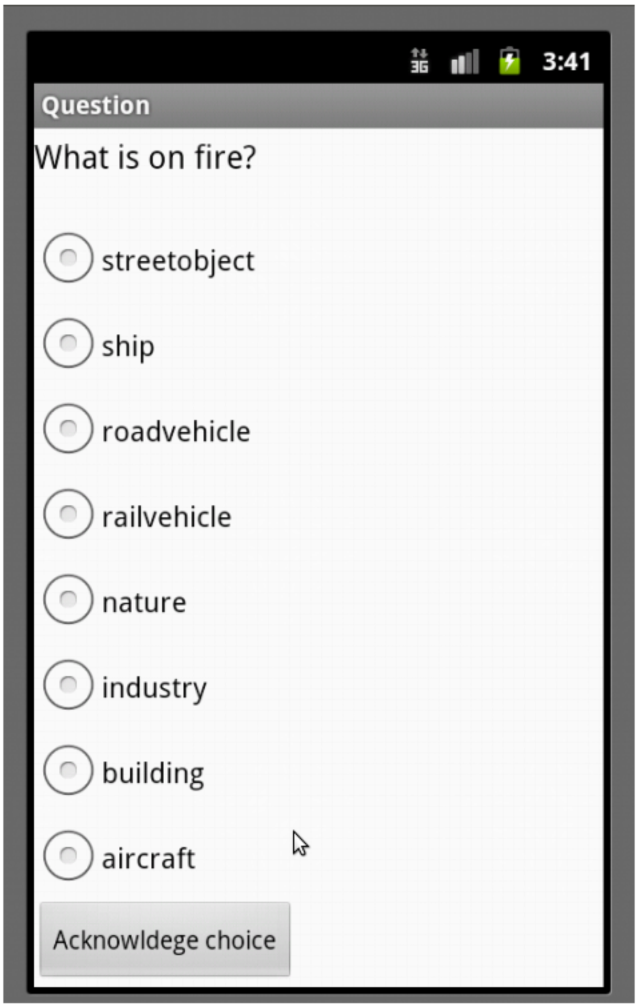
![ht] 

Figure 3.4: First screen generated from the ontology.

Algorithm 6 Algorithm used for SAQG.

```
Focus is set to FocalSituation
while Focus has subconcepts do
  SI is the set of supportedInfons in Focus
  if SI is empty then
    execute strategy 5
  else
    select strategy on the basis of SI
    execute strategy
    classify actual situation SA
    Focus is SA
  end if
end while
Focus is actual situation
```

abstract descriptions. Descriptions of both types have an informational value which is mutually comparable and indicates the semantic distance from the true and most specific description of the actual situation.

We developed several strategies to generate the most efficient sequence of questions to determine a situation. The choice of strategy depends on the assumptions that can be made about the domain. When situations are defined but no assumption about these situations can be made, a laborious strategy is used which subsequently asks whether a specific situation is true. Are situations or a part of a situation defined by parametric infons then we use a strategy which uses each parametric infon and asks whether one of the referents of the parameter is true. When a set of infons with unknown polarity is combined to create a series of possible situations then we use a 'divide and conquer' strategy. When that same assumption can be made and some of the theoretical possible situations are in practice impossible we use a strategy which prohibits impossible situations. When a situation is partially defined by a parametric infon and the set of references is defined in an ontology with is-a relations, we use a strategy which finds the most specific description.

With the Situation Awareness Question Generator (SAQG) we presented an application to automatically generate questions. These questions are generated from an ontology using several algorithms based on the strategies we developed. SAQG is meant to be used on a mobile device and gather information about a crisis situation from ordinary people. This information is then sent to emergency services.

Now we have the formalization and algorithms to ask the right questions at the right moment, we want to know next what kind of concepts we have to use. The next chapters will be about what the right concepts are and which strategies are optimal when SAQG is used by ordinary people for the purpose of determination of a situation.

Chapter 4

Determining Roles During a Crisis

Using the framework we created in Chapter 2, one of the strategies we elaborated in Chapter 3 and an ontology with concepts defined by experts we generate questions to determine the role a person can have during a crisis situation. An experiment is conducted to determine whether these questions are answered by ordinary people in a trustworthy way.

This chapter is based on a paper which was presented at SOTICS 2011: The First International Conference on Social Eco-Informatics (Teitsma et al., 2011a). The original paper is extended with an elaboration of the way in which we created the ontology and the analysis of the experimental results.

4.1 Introduction

In this chapter we investigate whether questions generated within the framework we created, comply with how ordinary people interpret such questions. The aim of the questions is to determine the role people have during the initial stage of a disaster. This role is used to characterize a person who is part of a situation.

The method we use to determine the different roles people can have during the response phase of a disaster, how these roles are related and how this influences the strategy, is described in Section 4.2. In Section 4.3 the experiment is presented. In Section 4.4 we discuss the results. Section 4.5 presents the conclusion.

4.2 Generating Questions to Determine Roles

4.2.1 Distinct Roles

Contrary to the idea of movie makers, politicians and the general public, people are not struck by panic when they are involved in a crisis situation. It has been found that, even during the most agonizing moments, people tend to help each other and can act rationally. Behaviour of people during a disaster is not very different from their normal behaviour. People tend to help other people, even if they do not know them (Clarke, 2002). Thus there are at least two groups of people during a disaster: active people who are willing to give a hand mitigating the effects of a disaster and passive people who need help or are not willing to help. Active people are people who are able and willing to help and observe (*Helper*) or people who only want to observe (*Observer*). Passive people are victims affected by the disaster in such a way that they need help (*Victim*) and people not physically affected but for some reason don't want to be active (*Not-Active*). Our classification then consists of the roles: *Helper*, *Observer*, *Victim* and *Not-Active*. All of these roles are subordinate to the concept *Person*.

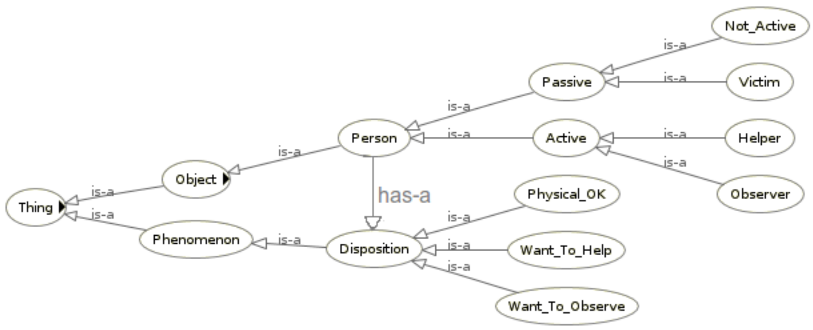


Figure 4.1: An ontology for roles during a disaster.

The roles we distinguish are differentiated by attributed properties. We use these properties to generate questions to determine the specific role for a person. Using properties to generate questions, the number of properties per role must be kept to a minimum to enhance efficiency. Furthermore, they shouldn't be ambiguous. The third requirement for the properties is that they must be based on the thoughts and feelings of the person him- or herself. Whether people want to help other people or not, depends on their disposition to help. The same subjective perspective must be applied to the willingness to observe and even the physical condition of the people we approach.

Figure 4.1 shows the resulting domain ontology. The concept *Disposition* is subordinate to *Phenomenon* and the concept *Person* is subordinate to *Object* in STOR (see Chapter 2). We distinguish three dispositions: *PhysicalOK*, *WantToHelp* and *WantToObserve*. Each *Person* has a *Disposition* which determines his or her role, i.e. *Helper*, *Observer*, *Victim* and *Not-Active*.

Infon	**Concept**	**Victim**	**Not-Active**	**Observer**	**Helper**
σ_1	Person	x	x	x	x
σ_2	Physically OK		x	x	x
σ_3	Want to Observe			x	x
σ_4	Want to Help				x

Table 4.1: Infons and concepts defining the four roles.

In Table 4.1 the properties of the roles are shown. Each property is presented as an infon (σ_n) which is the smallest possible informational entity (see Chapter 2). As can be seen in Figure 4.1 all the roles inherit from *Person*. All users of the system are persons. Our ontology consists of definitions of the form:

an observer is a *Person* who is *Physical_OK* and *Want_To_Observe* and not *Want_To_Help*. Each property is a piece of information we want to ask about. Such a property will be formulated as follows:

$$\langle\langle \text{hasDisposition}, IND_{person}, wantToObserve, \dot{l}, \dot{t}, \dot{p}\rangle\rangle \tag{4.1}$$

where IND_{person} refers to an actual person, $\dot{l}$, $\dot{t}$ and $\dot{p}$ are parameters for a specific location, time and polarity. Taken together, such compound infons can describe the role a person can have in a *Situation*. The roles we distinguish are a specification of the *Person* who is part of a *Situation*.

4.2.2 Dependencies

Trying to determine which situation is the actual situation, one easily creates an enormous amount of possible situations. The number of answers to a question determines how many situations are possible as description of the real situation. A "yes" or "no" as answer gives per question two possible situations and the addition of "I do not know" results in three possible situations. When having more than one question this results in a large number of possible situations. For example, 4 questions with each 3 possible answers gives 81 ($= 3^4$) possible situations. One has to constrain this combinatorial explosion.

Having four properties and restricting their value to two values gives 16 ($= 2^4$) possible situations, as can be seen in Table 3.1. Here S_{15} describes an *Observer* when σ_1 is the infon, which says someone is a *Person*, σ_2 describes that someone is *PhysicallyOK*, σ_3 that this person *wantToObserve*, σ_4 (*WantToHelp*) is not confirmed, stating this *Person* does not want to help. We then restrict the number of possibilities by determining dependencies between the properties.

There are three dependency relations in our ontology: the relation between 'being physically OK' and 'wanting to observe' and the relation between 'wanting to observe' and 'wanting to help'. Because of transitivity we can detect a third dependency between 'being physically OK' and 'wanting to help'.

This definition of concepts results in sets, which are subsets of other sets:

$$\begin{aligned} WantingtoHelp &\subseteq WantingtoObserve \\ &\subseteq PhysicallyOK \subseteq Person \end{aligned} \tag{4.2}$$

This equation says that the set of people who want to help is a subset of the people who want to observe, which is a subset of the people who are physically OK, which is a subset of persons. Here we see that when someone is physically OK it implies being a person. And when someone wants to observe it is implied he is physically OK. Summing up all the dependencies we come to the following list:

$$\begin{gathered} \sigma_2 \rightarrow \sigma_1 \\ \sigma_3 \rightarrow \sigma_2 \\ \sigma_3 \rightarrow \sigma_1 \\ \sigma_4 \rightarrow \sigma_3 \\ \sigma_4 \rightarrow \sigma_2 \\ \sigma_4 \rightarrow \sigma_1 \end{gathered} \tag{4.3}$$

What is stated here is that when in a situation A an infon σ_2 is supported then in situation A the infon σ_1 is also supported. The list also shows the most independent infon being σ_1.

In a system with logical dependencies, one should not expect that all the varieties given in Table 3.1 have an even chance of becoming real. It may even be so that some situations are impossible as outcome of a deliberation. The dependencies we formulated determine that situations in our system are possible or impossible. Whether a situation is possible or impossible is not known to the users of the system. The dependencies between the roles create a number of impossible combinations. How we keep users away from these impossible combinations is shown in the next section. First the impossible situations have to be determined.

	$\boldsymbol{\sigma_1}$	$\boldsymbol{\sigma_2}$	$\boldsymbol{\sigma_3}$	$\boldsymbol{\sigma_4}$
S_9	1	0	0	0
S_{13}	1	1	0	0
S_{15}	1	1	1	0
S_{16}	1	1	1	1

Table 4.2: Possible situations when taking dependencies into account.

The dependencies we have defined in the ontology restrict all the situations as mentioned in Table 3.1 to possible situations. Because all the roles are dependent on σ_1 this infon must necessarily be part of the situation. Looking at Table 3.1, it is obvious which situations are impossible: $S_1 ... S_8$. But also S_{10}, S_{11} and S_{12} are impossible, because in these situations people want to observe or help but are not physically OK. Lastly, S_{14} is impossible because this person wants to help but not observe, which we also ruled out. A summary of possible situations is shown in Table 4.2.

Knowing the properties of the roles and their dependencies it would appear to be the most efficient strategy to ask whether people want to observe. A question tree starting with σ_3 is shown in Figure 4.2. Within two questions the right role would be determined.

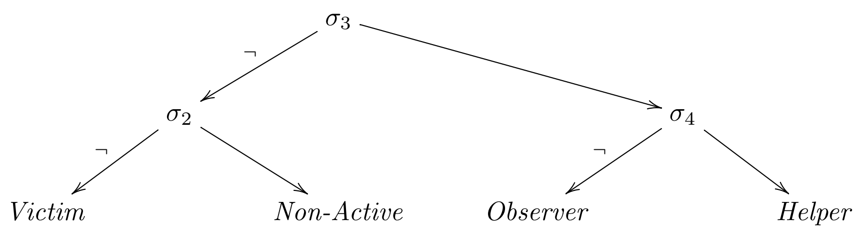


Figure 4.2: Decision graph starting with σ_3.

But then, when answering the first question (derived from σ_2), we suppose people know that answering 'yes' means they want to observe *and* are physically OK, which is an assumption we can not make. Because we can not make such assumptions, another method to determine the sequence of asking questions has to be found.

4.2.3 Semantic Strengthening

Now we know which situations are possible, we can determine which infon we have to ask for first. What we are after is an order of questioning which leads to the roles as defined in the ontology. The roles are defined by their properties which are represented as infons in the situations. Dependencies result in restricting the possible situations and excluding the impossible ones within our ontology. But these restrictions are not known by the persons that use our system. In this section we describe a method to preclude the impossible situations or prohibit the assignment of roles not in line with our definition of these roles.

Statements	**Number of referred situations**
σ_1	8
$\sigma_1 \wedge \sigma_2$	4
$\sigma_1 \wedge \sigma_2 \wedge \sigma_3$	2
$\sigma_1 \wedge \sigma_2 \wedge \sigma_3 \wedge \sigma_4$	1

Table 4.3: Classes of dependency.

We want to exclude the impossible situations as determined in Table 3.1. Including these situations as possible, yields semantic inaccurate information. Semantic vacuous statements, on the other hand, are true but not as precise as possible (see Chapter 3).

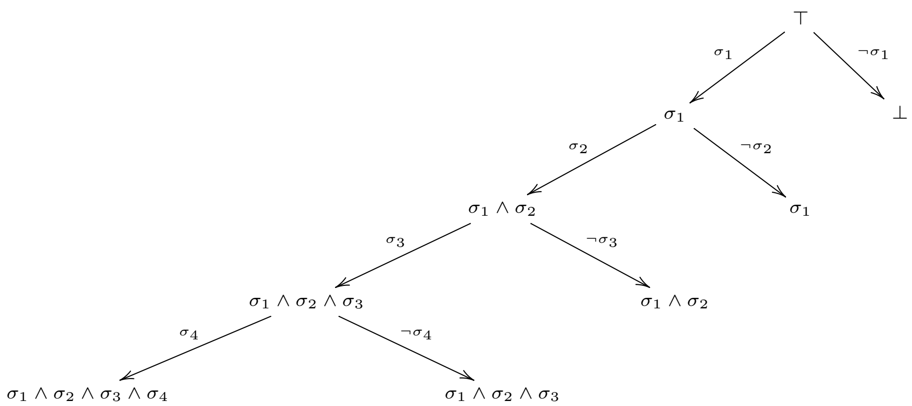


Figure 4.3: Decision tree.

The method we use, which we call *semantic strengthening* (see Section 3.3.4), is keeping the truthfulness when bypassing impossible situations. In our method we emphasise not the disjunctions but the conjunctions. Instead of replacing conjunctions with disjunctions, we work the other way around: replacing disjunctions with conjunctions. And the conjunctions are placed in such a way that there is no loss of truthfulness and impossibilities are ruled out.

Our roles are defined by the supporting infons which have a dependency as explained in Section 4.2.2. This makes some situations from the probability experiment (see Table 3.1) impossible. In our method we use these dependencies to find a sequence of questions and determine which question to ask first. As stated in Section 4.2.2 an optimal strategy would lead to asking after σ_3 (*WantToObserve*) as the first question. But this does not enforce the truthfulness

Confirmed infon(s)	**Situation**	**Follow up question**
σ_1	Person	Are you physically OK?
$\sigma_1 \wedge \sigma_2$	Physically OK person	Do you want to observe?
$\sigma_1 \wedge \sigma_2 \wedge \sigma_3$	Physically OK person who wants to observe	Do you want to help?
$\sigma_1 \wedge \sigma_2 \wedge \sigma_3 \wedge \sigma_4$	Physically OK person who wants to observe and help	No further questions asked.

Table 4.4: Mapping decision tree (see Figure 4.3) on question tree (see Figure 4.4).

because it jumps to conclusions. Our method starts with a truthful description of the situation and gradually, while describing the situation as it is, comes to the most specific description which is still true.

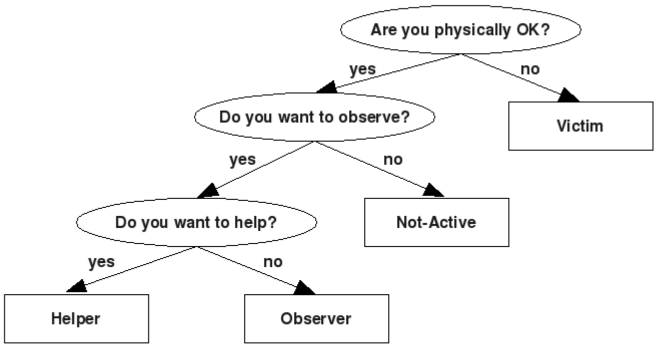


Figure 4.4: An example of a question tree.

The algorithm to determine the first infon to ask for is completed in two steps. First we determine which situations are possible and then we ask for the infon which is most common. Table 4.3 shows which infon should be used first to ask a question: σ_1. An assumption we make here is that when a mobile telephone is answered, this is done by a person. Thus the context suggests that σ_1 is true: we are asking a *Person*. The second question is about σ_2 and so on. By asking the questions in this order we prohibit impossible situations to occur.

In Figure 4.3 this order is also shown. Figure 4.3 starts at the top with the situation which is always true: the tautology, all situations as stated in Table 3.1 are true. When the call is answered it is inferred we are dealing with a *Person* and σ_1 is true. When it is determined that σ_1 is not the case then, in our ontology, this results in the situation which is always false: the contradiction. When σ_1 is true the second question is asked: 'Are you physically OK?'. When σ_2 is not the case it is determined that the situation is described sufficiently by σ_1 and the role *Victim* is assigned. When the user answers 'yes', he/she is physically OK and we know that $\sigma_1 \wedge \sigma_2$. The next question is 'Do you want to observe?'. When the answer is 'no', this *Person* is physically OK but does not want to observe and is called *Not-Active*. When the answer is 'yes', we know that $\sigma_1 \wedge \sigma_2 \wedge \sigma_3$ and the follow up question is 'Do you want to help?'. Answering 'no' makes the user an *Observer* and 'yes' a *Helper*. No further

Number	Actual role	Scenario
1	Victim	During the earthquake you were just drinking coffee in the kitchen. When you noticed the first trembles your ran out of the house but unfortunately a lot of debris was falling down and hit you. You have broken your leg and are not able to move. The telephone rings.
2	Not-Active	You woke up in the middle of night when a police car was riding down the street calling everybody out of bed and warning for an immediate flooding. The police warned not to flee but instead look for a high place and take food and drinks with you. You immediately went to the refrigerator took food and drinks and climbed through the bedroom window to the roof. But now you are sitting there and it is getting colder and darker. The street lights are not burning any more, probably because the power is down and you hear water streaming but see nothing. You are getting afraid and what is even worse you lost your glasses so you can't see clear. After a while the telephone rings.
3	Observer	After the first trembles you and your family ran out of your house. Luckily everybody came out of the house and now you are on the street. Your youngest child is only 3 months old and is sleeping now in your arms. Your 4 year old son is excited and wild probably because he is afraid. Your wife has quite a job to handle him. Your house has big cracks in it and you are afraid to go inside. Then the telephone rings.
4	Helper	During the earthquake you were walking in the park with your dog. You saw houses collapse and after five minutes when the earthquake seemed have come to an end you went for your house. But your house wasn't standing any more and collapsed like most of the houses in the street. Now you are in the street and the telephone rings.
5	Victim	Like every morning you go to your work by train. You are sitting next to a man who you slightly know because you see him most mornings taking the same train as you do. Suddenly you see a flash of light and then nothing is to be heard or seen for a couple of seconds, minutes or hours. When you recover from this blow you are lying on the floor and when wanting to move your leg you are feeling a lot of pain. When the telephone rings you can answer it.
6	Not-Active	You are sitting in a bus together with one of your friends. Suddenly the bus takes a turn and rolls over down a hill. All the people in the bus get shaken out of their seats and so are you and your friend. When all is coming to standstill you are going to look for your friend. Unfortunately you find him dead. It is hard to keep the tears from coming and on your knees next to your friend you start mourning. Then the telephone rings.
7	Observer	You and your family (your wife and three children) have been on the roof of your house for some hours now. It is cold, you have nothing left to eat and just a little water to drink. The water is still streaming hard and you see all different things floating around like cars, trees and even dead dogs. Then the telephone rings.
8	Helper	You work in a 20 story high building on the 14th floor. Just before the workday is over an alarm rings. Wondering what that is all about you go to the corridor and you smell something burning and see some smoke coming from the elevators. There seems to be a fire somewhere in the building. Then the telephone rings.

Table 4.5: Scenarios for the experiment.

questions are asked. The efficiency of the order of questioning is maximal, i.e. after each answer the total number of situations, as given by Table 3.1 is cut in half. Using the decision tree as shown in Figure 4.3 and the infons according to Table 4.4 results in a question tree as shown in Figure 4.4.

4.3 Experiment

The goal of the experiment was to find out whether human participants answered the questions posed in the same way as hypothesized by our theoretical framework and the representation of dispositions. To get familiar with and enhance the experimental method we first did two pilot experiments before doing the experiment as described here. In the first experiment we presented the questions as such. Because a lot of the participants chose a role which according to our ontology was not possible we presented the questions as shown in Fig. 4.5. In this diagram it is more clear how the answers determine a specific role. After the two experiments we had short interviews with some participants. The participants told us that some scenarios were not clearly stating the physical and mental disposition of the simulated persons. This feedback led to writing scenarios which were less ambiguous. We also learned that the questions and scenarios were difficult to understand because they were not stated in the mother language of the participants, i.e. Dutch. Accordingly, all the scenarios were translated in Dutch.

4.3.1 Experimental Methods

Different crisis scenarios such as an earthquake, a flooding or a bombing were used to describe a situation where people are involved in, immediately after the occurrence. For each scenario an actual role was envisaged, i.e. the roles participants should choose according to our hypothesis (see Table 4.5). For example, when a person was told 'you have a broken leg' (scenario 1), this person should take the role of victim according to our ontology. We introduced the questions beforehand and gave one example of the dependencies we had defined. Also, we used a flow diagram per scenario, as shown in Figure 4.5, to collect the answers for that scenario. With this flow diagram the participants were led to the appropriate questions and discouraged to give answers to questions which were excluded according to our ontology. Furthermore, we stated this instruction, the questions and scenarios in Dutch which is the native language of most of the participants.

Participants

In this experiment 38 students participated, all of them male and between the age of 18 and 22. All of the participants were students of the Amsterdam University of Applied Sciences.

Analysis

For the analysis of the data four metrics were computed: recall and precision, the Matthews Correlation Coefficient (MCC) for correlation (Matthews, 1975) and the F_1-score for accuracy (van Rijsbergen, 1974). MCC (also known as the ϕ-coefficient) is a computation of correlation between what is actual and what is predicted by a system or humans as in this case. Therefore so-called confusion matrices were needed to compute the values. Before we elaborate on how we use confusion matrices the general application of confusion matrices is explained.

In our experiment we showed people a particular situation in which we focused on one role, and wanted to know whether they could determine this situation as such. A confusion matrix represents the answers of the participants in relation to the actual role (see Figure 4.6). The actual role is the role we hypothesized as the role the participant chooses in a particular scenario. In our experiments we did not just show a positive and a negative example but

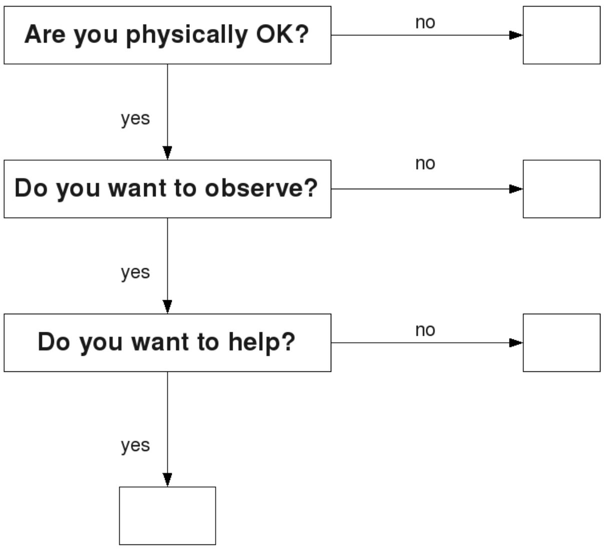


Figure 4.5: Flow diagram used in the experiment.

		Predicted value	
		Positive	Negative
Actual value	Positive example	True Positive	False Negative
	Negative example	False Positive	True Negative

Table 4.6: General use of a confusion matrix.

four distinctive examples. In constructing a confusion matrix for a particular role, only the scenarios with that role as the actual role were considered positive examples. All other scenarios were considered negative examples.

When the predicted role was the same as the actual role this still led to a ‘true positive’. When another role than the actual was predicted by the participant the outcome was a ‘false negative’ in the confusion matrix for *this actual role*. When that specific role was predicted instead of *another actual role* (in another scenario) this would lead to ‘false positive’. All other outcomes led to ‘true negative’.

An example may be helpful. Each scenario shown to the participants had an actual role which was envisaged e.g. *Victim*. When the participant answered the questions so that the result was that he was a *Victim*, this was marked as ’true positive’ in *this (first)* confusion matrix. When the participant chose for the role *Not-Active*, *Observer* or *Helper*, this was marked as ’false negative’. When another scenario was presented, with another actual role e.g. *Helper*, and, as a consequence of the answers, the role of the participant was determined as *Victim*, this was marked as ’false positive’ in the *first* confusion matrix. When the role was determined in that scenario as something other than *Victim*, this was marked as ’true negative’.

We used four metrics to interpret the results. MCC was used to tell whether there is a correlation between the actual and predicted values. It is a robust coefficient because it does not deviate when classes of different size are considered. MCC varies between -1 and +1 where -1 indicates a perfect negative correlation and +1 a perfect positive correlation, 0 indicates a random relation. The F_1-score is a metric which computes accuracy and varies between 0 and 1, where 0 indicates no accuracy at all and 1 a perfect accuracy. The F_1-score is the harmonic mean of recall and precision. The recall is a metric which computes how many of the actual situations are determined as such. Recall is computed as 'true positive' divided by the sum of 'true positive' and 'false negative'. Precision reflects how many of the predicted situations are actually these situations. It is being computed as 'true positive' divided by the sum of 'true positive' and 'false positive'. Precision is different from accuracy. While accuracy tells something about how close to the actual value the participants were, precision is about the variation of the predicted situations.

4.3.2 Results

The results can be seen in Table 4.7. The actual role was most predicted for *Victim* and *Helper*. A bias towards *Helper* can be seen when the participants were confronted with scenarios involving another actual role. The role *Not-Active* was the least clear, i.e. when the actual should have been *Not-Active* the participants chose mostly for other roles. The role *Observer* was slightly more clear.

	Predicted value			
Actual value	Victim	Not-Active	Observer	Helper
Victim	37	0	0	1
Not-Active	1	10	6	21
Observer	1	1	14	22
Helper	0	1	6	31

Table 4.7: Results of the experiment.

In Table 4.8 the analysis is shown. As can be seen there is a positive correlation for all the roles and for *Victim* even a strong correlation and accuracy. Recall and precision for *Victim* are also high which shows that for the participants it was clear when to choose this role. The role *Not-Active* was not often determined as a consequence of the answers, but when it was determined it was rightly determined. The role *Observer* was not often chosen (low recall) and when it was chosen often another role would have been more appropriate (low precision). The bias towards *Helper* can be seen in the high recall, i.e. often the chosen role was *Helper* and the low precision, i.e. often the role *Helper* was not the role envisaged by us. In our ontology (see Fig. 4.1) *Victim* and *Not-Active* are subclasses of *Passive* and *Observer* and *Helper* are subclasses of *Active*. When the roles are combined in *Passive* and *Active* the correlation is stronger. Also, the values for F_1, recall and precision for the combined roles are higher.

	MCC	F_1	Recall	Precision
Victim	0,95	0,96	0,97	0,95
Not-Active	0,39	0,40	0,26	0,83
Observer	0,26	0,42	0,37	0,48
Helper	0,30	0,51	0,74	0,39
Passive	0,63	0,76	0,63	0,94
Active	0,63	0,83	0,96	0,72

Table 4.8: MCC, F_1, recall and precision for the experiment.

4.4 Discussion

When we presented a scenario in which the participant was a *Victim* this was recognized by the participants and the role *Victim* was chosen. The constraints for other roles, especially the role *Helper*, were less clear. It seems the participants, when confronted with such crises, did only make a difference between victims and other people.

The combination of the roles *Victim* and *Not-Active* as *Passive* and *Observer* and *Helper* as *Active* yield higher associations with actual roles. This shows that our classification system, i.e. the questions we use to differentiate between the roles, is not precise enough for the roles we defined. The concepts *PhysicallyOK* and *wantToObserve* which we use as properties of the defined roles are not interpreted by the participants of our experiments in the way they are formally defined. The fine-grainedness of our ontology in which we define the roles, has to comply with the concepts people use. There is a difference between the formal definition of the concepts in the ontology and the semantic interpretation people have of these concepts.

4.5 Conclusion

In this chapter we have described an ontology we used to formulate questions. First we defined concepts in an ontology and then described dependencies between these concepts. With the help of a new method called *semantic strengthening* we found out which question to ask first. This method was used to prohibit choices which were precluded by our definitions of concepts. From the conducted experiment we concluded there is a relation between our ontology and the observations. The role *Victim* was clear for the participants but the other roles were not. When we combined the role as *Passive* and *Active* the relation between the actual active and passive roles and predicted active and passive roles was much stronger.

In spite of the improvements obtained after two pilot experiments, the system is still not capable of determining a role in all cases. In the experiment this relation was not strong enough to use it without restraint. Thus we conclude that the reasoning the system does on the basis of the answers of people, ought to be augmented by checking and confirming the answers provided. Using our framework with the ontology as created in this chapter, it will be necessary to ask further questions and probably to ask for confirmation after assigning a task. An example of such a question is ‘Can you move to the place where a Victim is?’.

Furthermore, it appears that the concepts we use are sometimes interpreted in another way than we meant. Apparently, people have a different mental model and reason in a different manner than represented in our ontology. This discrepancy results in a difference between actual role determination and our hypothesis. Therefore, we have to reconsider the construction of the domain ontology we use to generate questions. For this we have to look for concepts used by ordinary people, instead of concepts constructed by knowledge engineers.

Chapter 5

The Car Accident Experiment

In this chapter we study the question whether a situation can be determined by automatically generated questions consisting of commonly used concepts. The concepts we used in the previous experiment were defined by knowledge engineers. These concepts were interpreted by the participants in a different way then intended. This different interpretation led to answers which were not hypothesized. By asking question consisting of commonly used concepts we expect to enhance the comprehensibility of the questions and thereby the trustworthiness of the answers. Furthermore, instead of presenting participants a written scenario we show the participants a video and ask questions about visual objects. To generate questions we combine the second and third strategy as presented in Chapter 3.

This chapter is based on a paper presented at DEXA 2011: 22nd International Conference on Database and Expert Systems Applications (Teitsma et al., 2011b). The paper is extended with an elaboration of the analysis and results.

5.1 Introduction

In the previous chapter we concluded that the ontology we constructed was difficult to comprehend by ordinary people. In this chapter we investigate an alternative ontology for generating questions in actual situations. We construct such an ontology with commonly used concepts. We ask questions about situations which are presented in a video.

The representation we used in the previous chapter was defined by experts. This ontology did not generate questions leading to trustworthy answers to determine a situation. The concepts were interpreted by the participants of our experiments in a way we did not anticipate. Also, the situation we tried to determine, i.e. the situation of the user and his willingness to help, was inherently subjective.

To find concepts which are easily understood by people we look at concepts which are used by ordinary people. Such concepts can be found in systems which gather data from professionals such as police officers and ambulance staff. These professionals are experts in their own right but not experts in the field of representation and knowledge as knowledge engineers are. With these commonly used concepts we build an ontology and connect this to the Situation Awareness Ontology Revised as described in Chapter 2.

To test this ontology we conduct an experiment to find out whether the automatic generation of questions from such an ontology leads to a trustworthy determination of a situation. With our Situation Awareness Question Generator (see Chapter 3) we automatically generate questions from an ontology.

5.2 Method

To generate questions we use our Situation Awareness Question Generator (see Chapter 3). SAQG uses an ontology to compute which questions to ask. The domain ontology is constructed with concepts used in car accident databases and by domain experts. In the end SAQG produces yes-no questions and multiple choice questions.

5.2.1 Constructing the Ontology

To construct a domain ontology, we first determined what kind of concepts to use. For the generation of the ontology we used concepts which are as simple as possible. The reason for this was that we did not want people to necessarily make use of a great volume of background knowledge to answer the questions. Background knowledge of people often differs. Furthermore, consideration does take time and we wanted people to answer the questions as quickly as possible.

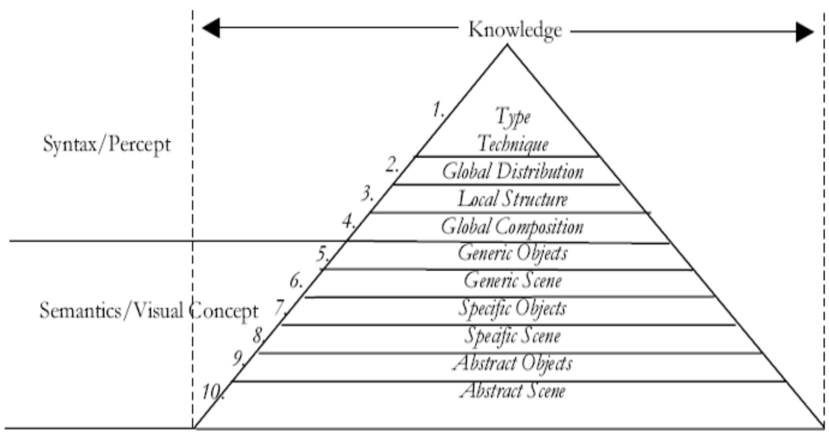


Figure 5.1: Pyramid of Jaimes and Chang (Jaimes and Chang, 2000).

The categorization by Jaimes and Chang (Jaimes and Chang, 2000) as shown in Figure 5.1 makes a distinction between percepts and visual concepts. Percepts are effects of nature and require little or no thought at all. A normal human being uses these concepts automatically. For example, reflected light carries information about shape, color, placement relative to other objects and other basic structures. Besides percepts Jaimes and Chang distinguish visual concepts which only exist due to knowledge and experience. For example, the perception of a rugby field which can not be comprehended or even recognized without knowledge of the (basic) rules of rugby. Percepts and visual concepts are classified into four syntactic and six semantic levels. Starting at level one, each subsequent level gets more knowledge-intensive. The use of concepts at the lowest levels (levels one to four) requires little knowledge and interpretation. Interpretation of concepts at the highest level on the other hand require the most knowledge. The concepts we used to construct the ontology and subsequent questions are in the categories 'Generic Objects' and 'Generic Scene'. Of the visual concepts these require the least knowledge. The interpretation of visual objects and their arrangement within these categories is solely based on everyday, common-sense knowledge. Examples of such objects are 'a passenger car', 'a van', 'rain', 'daylight'.

For the content of our ontology we used two sources. The first source for our ontology were two databases which gather all sort of data about car accidents. These data are extracted from reports created by emergency personnel such as ambulance staff and police officers. From all over the United States of America and the European Community such reports are accumulated. The databases we used are the Fatality Analysis Reporting System (FARS) (Administration, 2011) and Community database on Accidents on the Roads in Europe (CARE) (Commission, 2011). Many terms used as keywords in these database are concepts from the categories 'Generic Objects' and 'Generic Scene'. A concept such as *pavement* is immediately grasped by people and the subcategories *asphalt*, *concrete* and *unpaved* are easily understood.

The second source for our ontology was a traffic accident ontology based on domain expert knowledge (Yue et al., 2009). Six core concepts are identified, i.e. *time*, *location*, *weather*, *person involved in the accident*, *vehicle involved in the accident* and *accident event*. The last concept, which refers to the kind of accident (for example, *straight accident* and *overtaking accident*), was not used because it refers to (domain) expert knowledge.

Figure 5.2 shows how the key concepts in our ontology are related. *Car_accident* which is a subclass of *CrisisSituation* (see Chapter 2) has a *supports* relation with two infons: *MultipleVehicleInfon* and *InjuredInfon*. Also, *Car_accident* has a *hasAttributingInfon* relation with several concepts such as *Weatherinfon* and *Sort_Of_VehicleInfon*. Each infon has a parameter which refers to referents, varying in number from 4 to 8, which are subclasses of *Phenomenon* (see Chapter 2). The total number of classes in this domain ontology is 79.

Each infon is an informational entity, which makes it easy to formulate a question asking precisely for this piece of information. Each infon then has a question as a label. This label can be read when an infon is chosen as the most informative infon. When wanting to ask whether people are injured, the reading of the label will give 'Are people injured?'. The answer, 'yes' or 'no' equals the polarity of the parametric infon. The answers to the questions are kept in the ontology representing facts about the world.

5.2.2 Categories of Questions

The goal of SAQG is to determine a situation. A situation is defined with infons which are pieces of information denoting the actual situation as described in

Figure 5.2: Key concepts in the ontology.

the previous chapters. When asked for the actuality of these infons a yes-no question results. In the car accident ontology an abstract situation is defined by two parametric infons. This definition generates in SAQG two questions: 'Are people injured?' and 'Is more than one vehicle involved?'. The answers to these two questions are used to determine whether this or another situation is actual according to the user of the application.

When the situation is determined, additional information to obtain details about the situation is required. Such detailed information is elicited through parametric infons (see Chapter 2). An example of a parametric infon σ used in this ontology is:

$$\sigma = \langle\langle \text{Weather}, \text{is}, \dot{a}, 1\rangle\rangle \tag{5.1}$$

where $\dot{a}$ is a parameter of the type *Weathertype*. It is stated here that the weather is of a specific weathertype. It is important to realize that such an infon σ does not give any information about a situation when σ is one of the infons constituting that situation. To yield information, $\dot{a}$ first has to be anchored to some specific infon. For example, this set of weather types consists in our ontology of *dry*, *misty*, *hail*, *rain*, *snow* and *weather unknown*. These weather types are disjoint, i.e. two types of this set can not both be true. Most of the parametric infons in our application anchor their parameter in sets with disjoint classes. Only one parametric infon has a set with non-disjoint classes. This is the infon about the kind of vehicles involved in the accident. All of the answers to the questions are possible in some situation (Papasalouros et al., 2008). All the questions asked and the possible answers are shown in Table 5.1. Each question is of a category mentioned above.

Question	**Possible answers**	**Type of question**
Are people wounded?	Yes No	yes-no question
Is more than one vehicle involved in the accident?	Yes No	yes-no question

How are the lighting conditions?	Daylight Night with artificial light Night without artificial light Dawn or dusk Lighting conditions unknown	multiple-choice-question with one possible answer
How are weather conditions?	Dry Fog Hail Rain Snow Unknown	multiple-choice-question with one possible answer
On what kind of road did the accident happen?	Free way Within an urban environment Regional road Highway Unknown	multiple-choice-question with one possible answer
How is your line of sight?	Severly limited Limited Unlimited Unknown	multiple-choice-question with one possible answer
What is the condition of the road?	Dry Wet Snowy Glazed frost Unknown	multiple-choice-question with one possible answer
What kind of pavement does the road have?	Asphalt Concrete Cobbels Unpaved Unknown	multiple-choice-question with one possible answer
How many lanes are there, in one direction?	1 lane 2 lanes 3 lanes 4 lanes 5 lanes More than 5 Unknown	multiple-choice-question with one possible answer
How many vehicles are involved in the accident?	2 vehicles 3 vehicles 4 vehicles More than 4 vehicles Unknown	multiple-choice-question with one possible answer
What sort of vehicles are involved in the accident?	Van Moped Bicycle Car Tractor Truck Unknown	multiple-choice-question with more than one possible answer

Table 5.1: Questions for the experiment.

Not all questions were asked each time the application was used. Several constraints were built-in which made some questions redundant because by implication the answer was already known. An example of such a constraint is 'when the weather type is rain then the street is wet'. When the participant answered *rain* to the question what kind of weather it was then the question after the condition of the street was not asked. These constraints minimized the number of questions to be asked. The constraints are shown in Table 5.2.

if		**then**
weather condition	is *rain*	do not ask for the condition of the pavement
	is *hail*	
	is *snow*	
	is *fog*	only offer ‘Severely restricted and ‘Restricted’ as possible answer for the question after the line of sight
kind of road	is *within an urban environment*	do not ask for the number of lanes
	is *regional*	
	is *unknown*	
	is *freeway*	do not ask for the kind of pavement
	is *highway*	
situation	is *With_injury_and_one_vehicle*	do not ask for the specific number of vehicles involved in the accident
	is *Without_injury_and_one_vehicle*	
	is *With_injury_and_multiple_vehicles*	do not offer ‘1 vehicle’ as answer for the question after the number of vehicles involved in the accident
	is *Without_injury_and_multiple_vehicles*	

Table 5.2: Constraints used in the experiment.

5.3 Experiment

An experiment was conducted to find out whether questions without the use of knowledge-intensive concepts resulted in trustworthy answers.

5.3.1 Participants

The participants of the experiment were undergraduate students of the Amsterdam University of Applied Sciences. The 89 participants are characterized by being mostly male (78,65%) and between 18 and 22 (76,41%).

5.3.2 Material

We used the following videos:

- Video 1: *With injury and multiple vehicles.* It is night but there are street lights. The motorway has three lanes (oneway). The weather is dry. A car hits a van. Two people in the car are clearly in trouble (injured). Because the van is being pushed onto another lane it is hit by another car, starting a cascade of crashes (but no more injured people are shown).

- Video 2: *With injury and one vehicle.* In this video the perspective is from the passengers seat in a car on a motorway (three lanes) during daylight in dry weather conditions. Although the car is on the left side lane another car is trying to overtake it with great speed. The overtaking car gets of the road, hits the guardrail, smashes into a ditch and lifts of, hitting the pillar of a viaduct. No other car is hit.

- Video 3: *Without injury and one vehicle.* From the passengers perspective it is shown that there has been an accident on a motorway of three lanes and surrounded by a fence. Four people are standing on the side of the road, seemingly not hurt. A track of rubber ends at a damaged car.

- Video 4: *Without injury and multiple vehicles.* From a balcony (fourth or fifth floor) a video is shot of a snowy street with parked cars. A car rides down the street but can't stop because of icy road conditions. It hits several cars. A second car does the same.

The participants were asked to watch one of these videos which lasted between 27 and 37 seconds. First a short introduction was given and the goal of the experiment was explained. Each participant was told that there is the possibility that the video will be shocking to a small degree and when the participant does not feel comfortable any more it is no problem to stop. None of the participants stopped. When the participant had watched the video, questions generated by SAQG were presented on a mobile device.

5.3.3 Results and Analysis

For the analysis of the data the same four metrics as in the previous experiments (see Chapter 4) were computed: the Matthews Correlation Coefficient (MCC) for correlation (Matthews, 1975) and the F_1-score for accuracy (van Rijsbergen, 1974), recall and precision. For this we again used confusion matrices. The use of confusion matrices we created was slightly different from the use of confusion matrices in the previous experiments because participants could only choose yes or no and not, as in Chapter 4, four different roles. The use of confusion matrices is then much more straightforward.

Of the four videos shown, two were with injured people and two were without injured people. These were respectively positive and negative examples. When for example video 1: *With injury and multiple vehicles* was shown and the participant answered 'yes' when asked whether people were injured, this was marked as 'true positive' in *this* confusion matrix. When the participant answered 'no' to this question it was marked as being 'false negative'. Showing another video, for example video 3: *Without injury and one vehicle*, it was possible to answer 'yes' to the same question. Such an answer was being marked as 'false positive' for *this* confusion matrix. Seeing video 3: *Without injury and one vehicle* and answering 'no' to the question 'Are people injured?' marked a 'true negative'.

As described in Section 5.2, the answers to the two questions 'Are people injured?' and 'Is more than one vehicle involved?' were used to compute the determination of a situation. In the ontology, the abstract situation *Car accident* is supporting two parametric infons which both can have a polarity '0' or '1'. *Car accident* is thus an abstraction of four situations. When a particular video was shown, two yes-no questions were asked and people answered accordingly. Both answers were then used to compute the situation.

SAQG poses multiple choice questions after the initial situation was determined. The set of actual values is a subset of the possible answers to the question. When, for example, asked after the weather type, participants are offered a set of answers consisting of *dry*, *misty*, *hail*, *rain*, *snow*, *weather unknown.* The four videos shown to the participants reflected two different weather types: *dry* and *snow.* These two actual situations generated two confusion matrices although the number of possible answers was six. When the weather type in the video was *dry* and the participant answered something different this would be marked as 'false negative'. When the weather type was *snow* and the participant answered *dry* this was marked as 'false positive'.

Table 5.3 shows which situation was determined after the first two questions had been answered. We first computed whether there was a dependency between the actual and predicted values. This appears to be the case: χ^2 (9,N = 89) = 143,61, p<0.05 for H_0. The sum of all the right predictions ('true positives') in Table 5.3 shows that in $\left(\frac{19+19+16+16}{89} = \right)$ 78, 85% of all the cases the participants predicted the right situation.

	Inferred values				
Actual values	Situation 1	Situation 2	Situation 3	Situation 4	Sum
Situation 1: With injury and multiple vehicles	19	0	0	3	22
Situation 2: With injury and one vehicle	0	19	0	4	23
Situation 3: Without injury and one vehicle	1	2	16	3	22
Situation 4: Without injury and multiple vehicles	0	5	1	16	22
Total	20	30	17	22	89

Table 5.3: Determined situations inferred from answers given by participants.

Table 5.4 shows the MCC, F_1-score, recall and precision for the determination of situations and the yes-no questions used to determine these. The correlation between the actual values and inferred values was rather high, just as the accuracy. These values were the highest when the situation 'With injury and multiple vehicles' was shown. This video explicitly showed injured people and an accident with several cars involved. The situation 'With injury and one vehicle' was not so explicit, i.e. no one was seen to be injured and the participant had to infer by the severity of the accident that the driver and other people who might be in the car should be injured. This inference mechanism was also working the other way around when no one was injured and participants inferred the driver of the car having an accident should be injured because they interpreted the accident as being a severe accident. Another problem confronting the participants was how to interpret the concept of 'being involved' as in 'is there more than one car involved in the accident?'. On the one hand participants interpreted 'being involved' as 'being in the neighbourhood' or 'is part of what can be seen'. On the other hand it was being interpreted as 'is the cause of the accident'. This broad interpretation and the inference from the severity of the accident to injury of the driver lead to a great number of false positives in the confusion matrix of the situation 'With injury and one vehicle'. Hence the relative low MCC value for the situations 'With injury and one vehicle' and 'Without injury and multiple vehicles'.

In Table 5.4 it can be seen there is a difference between the mean of all the values for the questions and the mean of all the values for the inferred situations. The inference of situations thus goes with a (small) loss of information.

The values from the confusion matrices for the multiple choice questions

Actual infons and situations	**MCC**	$\boldsymbol{F_1}$	**Recall**	**Precision**
People are injured	0.76	0.88	0.93	0.84
Only one vehicle is involved	0.79	0.88	0.80	0.97
Mean of yes-no questions	0.78	0.88	0.87	0.91
With injury and multiple vehicles	0.85	0.89	0.83	0.95
With injury and one vehicle	0.64	0.73	0.86	0.63
Without injury and one vehicle	0.78	0.82	0.73	0.94
Without injury and multiple vehicles	0.64	0.73	0.73	0.73
Mean of situations	0.73	0.79	0.79	0.81

Table 5.4: Matthews Correlation Coefficient, F_1-score, recall and precision for the actual infons and situations.

are shown in Table 5.5. We have detected a number of questions to which the participants gave relatively less trustworthy answers. The highest recall value of the seven questions for which the participants gave a less trustworthy answer (marked in Table 5.5 by $*$) is 0,59 and the lowest value of the other questions is 0,72. The questions to which the participants gave a less trustworthy answer are characterized as being knowledge-intensive.

Confusion matrix	**MCC**	$\boldsymbol{F_1}$	**Recall**	**Precision**
Weather				
The weather is dry	0.73	0.90	0.83	0.98
The weather is dry (without video 1)	0.84	0.94	0.91	0.98
The weather is snowy	0.88	0.91	0.91	0.91
Road				
The road is a highway	0.82	0.95	0.90	1.00
The road is in town	0.91	0.93	0.86	1.00
The road is dry	0.67	0.84	0.72	1.00
The road is dry (without video 1)	0.94	0.97	0.95	1.00
The road is covered with snow	0.88	0.91	0.88	0.96
The road has three lanes	0.44	0.69	0.52*	1.00
The road has one lane	0.91	0.93	0.96	0.91
The pavement is asphalt	0.59	0.91	1.00	0.84
The pavement can not be known	0.59	0.58	0.41*	1.00
Visibility				
Acc. during daylight	0.82	0.91	0.89	0.93
Acc. during the night in artificial light	0.85	0.70	0.59*	0.87
It is dawn or dusk	0.48	0.58	0.50*	0.69
Sight is unlimited	0.22	0.65	0.73	0.59
Sight is limited	0.32	0.55	0.48*	0.64
Number of vehicles				
Passenger cars and vans are involved	0.52	0.62	0.55*	0.71
Passenger cars only	0.57	0.89	0.87	0.91
More than four vehicles are involved	0.62	0.73	0.59*	0.96

Table 5.5: Summary of all the MCC's, F_1-scores, recall and precision (* indicates a value smaller than 0,6).

When participants had to count, their observation became less trustworthy. Two answers, one about the number of cars involved and another about the number of lanes, scored significantly less correct than the answers where the participants did not have to count or where the counting was simple, e.g. *The road has three lanes* and *More than four vehicles are involved* vs *The road has* *one lane* and *One vehicle is involved.* Both answers were given to multiple choice questions where other answers than *one lane* and *three lanes* were possible. A lot of participants used these other answers as their observation leading to a low recall (a lot of 'false negative'), high precision (a few 'false positive') and an overall low correlation between the actual and predicted values. The answers *The road has one lane* and *One vehicle is involved* both show a high recall and precision. This led to a high correlation for the *one lane* answer, indicating that such an observation was easy to make. Restrictions on the observation about *One vehicle is involved* are given above.

When participants had to infer information it became more difficult to give the right answers. Asked about the sort of pavement in video 4: *Without injury and multiple vehicles* participants gave as answer *asphalt* while the question clearly could not be answered because the road was covered with snow. The high recall for *The pavement is asphalt* contrary to the low recall and high precision for *The pavement can not be known* even suggests a bias to *The pavement is asphalt* as the answer, when participants did not know what to answer.

An observation about the limits of sight was hard to make for the participants. Night or dawn as in video 1: *With injury and multiple vehicles* and video 4: *Without injury and multiple vehicles* gave participants problems to tell whether the sight was limited or not. A fence, as in video 3: *Without injury and one vehicle*, gave the same problems. Of all the MCC values the one for *Sight is unlimited* was the lowest and for *Sight is limited* the second lowest.

Doing observations during the night has proven to be difficult. Table 5.5 shows the impact of removing video 1: *With injury and multiple vehicles* for the questions about the weather type and whether the road was dry or not. In both cases the recall improves and precision stays the same, indicating the value for 'false negative' decreases. Nightly observations about the weather and road led to more wrong answers than day time observations.

When the mean of the time needed to answer the questions was taken of the different categories of questions, a difference is seen for the yes-no questions for which the mean was 6,86 seconds with a standard deviation of 4,67 and the multiple choice questions with one answer for which the mean was 9,34 seconds with a standard deviation of 5,54. For the multiple choice questions with more possible answers the time was 17,55 seconds with a standard deviation of 10,25. The mean of the time needed for the questions which resulted in less trustworthy answers as indicated by '*' in Table 5.5 (in Table 5.6 also indicated with '*') was 9,85 seconds with a standard deviation of 6,20. The mean of the other multiple choice questions was 8,45 seconds with a standard deviation of 4,43.

5.4 Discussion

The experiment showed that questions and presented answers constructed from commonly used concepts can be answered in a trustworthy manner. It also showed that when people have to *interpret* concepts used in questions the trustworthiness of the answers is lower. This corresponds to the theoretical framework of Jaimes and Chang (Jaimes and Chang, 2000). The counting of objects (vehicles or lanes) is difficult and leads to a drop in trustworthiness of answers. A question as 'What is the sight?' with possible answers *limited* and *unlimited* had to be interpreted. What we hoped to determine was whether mist or dark-

Question	Mean
Are people wounded?	6.71(4.50)
Is more than one vehicle involved in the accident?	6.70(4.40)
How are the lighting conditions?	9.83*(8.88)
How are weather conditions?	8.17(4.15)
On what kind of road did the accident happen?	8.83(4.27)
How is your line of sight?	9.81*(5.69)
What is the condition of the road?	8.28(4.88)
What kind of pavement does the road have?	8.60*(4.47)
How many lanes are there, in one direction?	11.12*(7.72)
How many vehicles are involved in the accident?	9.87*(4.22)
What kind of vehicles are involved in the accident?	17.55(10.25)

Table 5.6: Mean and standard deviation of the time needed per question ('*' indicates a question which elicits less trustworthy answers (see Table 5.5)).

ness limited sight but according to some participants a fence also limited sight. The concept of sight was thus not as unambiguous as we thought. Direct observations of visual concepts are trustworthy while indirect observations which require interpretation or inference are less trustworthy.

This experiment deviates from a real life situation in that the participants saw a video and were not participating in an actual car accident situation. Corresponding to a real-life car accident situation, people have time for reflection and answering questions *after* the rapid course of events. A difference from a real-life situation was that the observation we asked about, had to be remembered. When asked about the weather, in real life one looks (again) at the sky and answers the question. In our experiment the participants had to think back to what they had seen and try to remember. But this restriction is general and applies to all questions and thus makes no difference between the relative trustworthiness. The yes-no questions may be exceptions to this rule because in real life these are also not answered at the moment the accident occurs.

5.5 Conclusion

First we conclude that it is possible to automatically generate questions from an ontology about a situation combining the second and third strategy as presented in Chapter 3. When used by our Situation Awareness Question Generator such an ontology has to comply with the Situation Theory Ontology Revised as described in Chapter 2. When it does, automatically generated questions about several situations become available.

The second conclusion we draw from this experiment is that it is possible to gather trustworthy information from people who saw a hectic situation like a car accident. It should be noted though that there are certain restrictions such as when such an event is happening at night one should be more careful with this information than in broad daylight. All of the answers have a certain margin of truthfulness, i.e. the accuracy of the answers is not perfect. One has to be aware of multiple possible interpretations of the words such as 'sight' used in a question.

In this chapter we have found concepts of the category 'Generic Objects' and 'Generic Scenes' in systems used to gather data about car accidents. We used these concepts because of their ease of understanding. To create ontologies of concepts which are easy to understand we have to determine more specifically what characterizes such ontologies and how to create an ontology with these characteristics.

Chapter 6

Engineering Ontologies for Question Generation

Using an ontology to automatically generate questions for ordinary people requires a structure and concepts compliant with human categorization. Here we present methods to develop a pragmatic-based, an expert-based and a basic-level ontology and a framework to evaluate these ontologies. Compliance with human categorization was measured by using existing methods and measuring semantic distance of concepts in a new way. Comparing these ontologies shows that expert-based ontologies are the most easy to construct but lack required cognitive semantic characteristics. Basic-level ontologies have structure and concepts which are better in terms of cognitive semantics but are most expensive to construct. The pragmatic ontology is easy to construct but has some characteristics which makes it less suitable than the basic-level ontology.

This chapter was published as an article in Applied Ontology (Teitsma et al., 2014a). To make it consistent with other chapters a few adjustments were made and some syntactical refinements were done.

6.1 Introduction

In this chapter we investigate the construction and evaluation of ontologies for driving question generation in the field of crisis management. The basic idea is that in a crisis situation ordinary people involved in the crisis can be asked a series of questions to determine the details of this situation. We aim to contribute methods for ontology construction for this specific task using a combination of known and new techniques. In this chapter we describe the development of three ontologies to support the task of information gathering in the domain of crisis management. To evaluate these ontologies we developed a framework for which several known metrics are reinterpreted to determine suitability for the task. Furthermore, we extended the evaluation of the cognitive semantics of an ontology, i.e. the compliance with human categorization: not only completeness will be measured, but also entropy, cognitive ergonomics and a new metric called *semantic distance validation*. This framework gave us metrics to evaluate the suitability of an ontology for the specific task of generating questions. We conducted two types of experiments: to construct ontologies and to validate ontologies. In further research the three ontologies will be tested in a practical context using SAQG to elicit information from people watching a crisis situation (see Chapter 7).

The wide distribution of mobile devices makes it possible to involve a large group of people in an automated information gathering process to quickly assess a crisis situation. Through their mobile device humans can report on the situation they find themselves in. Our application uses concepts and their relations from an ontology to generate questions and possible answers. For example, when a car is on fire, the user is asked 'What is on fire?'. Possible answers are presented: 'Building', 'Aircraft', 'Roadvehicle', etc. The user chooses an answer: 'Roadvehicle'. The subordinate concepts of 'Roadvehicle' are presented:'Truck', 'Motor', 'Car', etc. The user chooses an answer again ('Car'). Answers to such questions steadily refine the description of a situation. In this chapter we focus on the subsumption relation to generate questions which gradually elicit more detailed information about specific objects from ordinary people. To get reliable and valid answers, the questions put to humans need to pertain to the situation at hand, be easily interpretable and have high discriminatory value. Because we use a domain ontology to automatically derive questions, the ontology should support these requirements. In addition the construction process should be efficient.

The experiment as presented in Chapter 4 revealed that an ontology solely based on the view of domain experts does not comply with human thinking and is not very suitable for the purpose of asking questions and gathering information from ordinary people in a valid way. Analysis of the data taught us that participants interpreted the concepts differently than intended by the domain experts. For example, domain experts use stricter definitions of concepts than ordinary people do. To overcome this problem, an ontology was constructed from concepts used by professional workers such as police officers and ambulance staff. Professional workers use concepts which are less detailed than the concepts domain experts use. With this ontology we generated questions that were answered in a more valid way than the questions generated in the experiment as presented in Chapter 5. However, these studies focused on the automatic generation of questions and the evaluation of the appropriateness of the answers provided. A principled investigation of the nature and structure of different ontologies used for the generation of questions was not part of these studies.

In this chapter we present three different types of ontologies for the domain of fire situations and different ways in which these ontologies can be constructed. We discuss the following three types of ontologies: *a*) an ontology developed from a categorization used at an emergency call center, *b*) an ontology constructed from existing ontologies created by domain experts and *c*) an ontology based on empirical data elicited from ordinary people. In addition we develop a framework of four classes of evaluation criteria to assess the value of the different types of ontologies for the task of automatically generating questions: *a*) the ontology must have a structure which is useful for the task at hand, i.e. question generation on a mobile device, *b*) the construction of the ontology is efficient, *c*) the ontology must be complete, i.e. all concepts that are relevant should be contained in the ontology and *d*) the ontology should be compliant with human categorization.

The remainder of the paper is structured as follows. The next section discusses background theory and related work. Section 6.3 presents the construction of an ontology from an emergency call categorization, an expert knowledge based ontology and an ontology created with concepts elicited from ordinary people. An evaluation framework is described in Section 6.4. An evaluation of the constructed ontologies using the evaluation framework is presented in Section 6.5. We discuss the results of this evaluation in Section 6.6 and end with our conclusions in Section 6.7.

6.2 Background and Related Work

In this section we present the background for constructing the three types of ontology. We conclude with a short discussion about an evaluation framework for these ontologies and the task they are used for.

6.2.1 Pragmatic-based Ontology

An obvious method to create a domain ontology is looking for vocabularies and simple categorizations used to perform a specific task in the domain. The value of such categorizations for professional workers is undeniable and the origin and evaluation of such categorizations gives them a pragmatic quality. But improvements are necessary to make them useful for the task of automatically generating questions. These categorizations often are constructed without any formalization and their structure is more or less heterogeneous, i.e. the concepts and their categorization do not have a strict definition and structure.

6.2.2 Expert-based Ontology

Traditional approaches to ontology construction rely on knowledge represented by collections of texts (Buitelaar et al., 2005) or knowledge from experts. Eliciting the knowledge from experts is done by a knowledge engineer who uses knowledge elicitation techniques which result in semi-formal descriptions. In a knowledge modelling process these are converted into a more formal description (Schreiber, 2008). Such ontologies are extensive and not developed with the goal of being comprehensible to ordinary people. The classical way to categorize concepts or objects of interest is by using crisp sets and fixed boundaries and is based on set theory. Concepts are defined in a strict and rigorous manner (Murphy, 2004; Frixione and Lieto, 2011; Taylor, 2003). This way of categorizing expert knowledge then seems to be the natural approach for building ontologies as used in knowledge engineering. A complete definition of concepts is preferable, i.e. it should state necessary and sufficient conditions (Gruber, 1995). The development and use of an expert-based ontology is not without problems. Such an ontology has to overcome the lack of precision of natural language when used in describing the world (Guarino and Welty, 2002). Expert-based concepts are created by domain experts such as scientists, bureaucrats and lawyers. The concepts used in the ontology do not automatically correspond to every day use of concepts by ordinary people. This discrepancy between formal reasoning and common-sense knowledge is, within the context of ontologies, far from being exhaustively studied (Fensel, 2004). Experts have knowledge of a domain which other people, i.e. laymen, do not have. Experts more easily make the distinction between relevant and irrelevant parts of the information they perceive. Inferences from information are made easier and faster by experts than by non-experts (Kellman and Garrigan, 2009).

6.2.3 Basic-level Ontology

To construct an ontology corresponding to human thinking we studied experiments in psychology describing the phenomena of basic-level concepts and prototype effects. Basic-level concepts are concepts at a particular level within a hierarchy of concepts which people prefer to use. The prototype effect is the phenomenon that some concepts are, within a group of subordinate concepts, more typical than others. These phenomena can not be explained with the classical theory of categorization which assumes an objective categorization of concepts solely based on their properties (Lakoff, 1987). Basic-level concepts can be found in human cognition and communication and descriptions of behaviour, events and scenes (Rosch et al., 1976a,b; Tversky and Hemenway, 1983; Murphy and Smith, 1982; Rifkin, 1985; Morris and Murphy, 1990). People are faster at verifying that an object is a member of a basic-level concept than they are at verifying concept membership for either subordinate or superordinate concepts (Murphy, 2004). Categorization of objects is based on their attributes as in expert-based ontologies. But attributes of concepts are not primitives, they are related to the use of the object, i.e. the function of an object and the way people interact with the object are taken into account when categorizing an object. This leads to a variety of categorizations in different cultures and times (Taylor, 2003).

Basic-level concepts can be found due to the characteristic that they have a stronger association with their attributes than concepts at their superordinate and subordinate level. People prefer using a basic-level concept to denote an object because it is the most economic trade-off between being too general and too specific. A general concept does tell not very much about a specific object because a large number of objects have the same attributes. A specific concept necessitates one to look very carefully because at such a level the differences between concepts become small, i.e. only a few or even just one attribute makes the difference. A basic-level concept improves the ability to predict the attributes of objects denoted by that concept. The extent to which a level of abstraction has this characteristic is called the 'category utility' of that level[1](Corter and Gluck, 1992).

Chen et al. (Chen et al., 2010) show that by using folksonomy tags for creating an ontology based on basic-level concepts the resulting ontology agrees more closely with human thinking than an ontology that was automatically generated from text. Macías-Galindo et al. (Macías-Galindo et al., 2011) also use tags from folksonomies to construct a categorization. They conclude that the quality of such a categorization can be enhanced by using a backbone: a set of root domain concepts provided by domain experts instead of using just one root concept.

[1] In psychology the term 'category' is used to denote what we here call 'concept'.

6.2.4 Ontology Evaluation

The characteristics of an ontology influence the degree to which questions and their proposed answers are useful to determine the situation. Most metrics of ontologies evaluate the structure of the ontology instead of its semantics (García et al., 2010; Ma et al., 2010). A framework for a broader evaluation of different types of ontologies is hard to find. Gangemi (Gangemi et al., 2005) offers a framework which identifies metrics along three dimensions of an ontology: the structural, functional and usability-related dimension. Structure-related metrics focus on hierarchical relations between concepts. The functional dimension is captured by measuring the discrepancy between the comprehension of a conceptualization (cognitive semantics) and its formal description (formal semantics) which is called the matching problem. The usability-related dimension, as defined by Gangemi, refers to the level of annotation of an ontology and gives a potential user meta-data about the ontology to compare it with other ontologies. Sabou (Sabou and Fernandez, 2012) presents a method to evaluate existing ontologies on four topics: domain coverage, quality of modelling, suitability for an application or task and adoption.

We develop the ontologies with the goal of using them for a specific task: generating questions to determine a situation. Metrics which are identified by Gangemi as showing the structural dimension of an ontology, i.e. topological and logical properties, are also useful to evaluate the functionality of an ontology for this task. Furthermore, of great importance for this task is the comprehensibility of the ontology by ordinary people. When the cognitive semantics of an ontology is evaluated it is usually restricted to completeness and frequency of use. The metrics for the matching problem suggested by Gangemi are all focused on the lexical level, i.e. these metrics show properties of singular concepts. In our opinion the matching problem is much broader and should also refer to cognitive ergonomics, perception of relations between concepts and the total number of choices offered by the ontology to get an indication of the compliance with human categorization. Therefore, we interpreted some known metrics in a new way.

Structure	**Completeness**	**Cognitive Semantics**
Number of concepts	Coverage	Entropy
Maximum path length	Precision	Ingve-Miller number
Average path length		Semantic distance validation
Coherence		
Maximum number of subclasses		
Average number of subclasses		

Table 6.1: Ontology metrics.

Our evaluation framework consists of metrics concerning structure, completeness and cognitive semantics of an ontology as shown in Table 6.1. We distinguish completeness from cognitive semantic metrics because completeness refers to the concepts themselves and not their relationships as do the cognitive semantic metrics. Cognitive Semantics refers to the human conception of the relations between concepts within a knowledge structure such as an ontology.

The metrics in Table 6.1 can all be used in different ways. First, as metrics of an ontology as such, i.e. a general evaluation of an ontology. Secondly, these metrics can also be seen as metrics of the comprehensibility of an ontology for people. For example, the number of concepts gives an indication of how easy (or difficult) it is for someone to comprehend the ontology. A third perspective on these metrics takes the specific task for which an ontology is used into account. In our research all these metrics are focused on the task at hand, i.e. generating questions to determine a situation.

6.3 Construction of Ontologies

To construct the three ontologies we first created a backbone as is shown in Section 6.3.1. During the construction of the ontologies we repeatedly made use of frequency lists. How we used these frequency lists and some points of interest regarding their use is presented in Section 6.3.2. The concepts we used to construct the ontologies are from the domain of fire situations. This domain is populated with concepts related to fire and everything that can be on fire.

6.3.1 Backbone

To construct the ontologies we made use of a backbone. This backbone is a set of domain root concepts and was based on an interpretation of the categorization as used by fire departments in the Netherlands. The domain concepts in the classification used by fire departments are:

- Outside. These are artifacts on the street like containers, bus shelters or waste-buckets.
- Building. Buildings such as a house, office building or shop.
- Industry. These are places for industrial activity.
- Aviation. Vehicles which are used to fly, such as an airplane or zeppelin.
- Nature. Areas consisting of forest, dunes, brush.
- Shipping. Vehicles which float on the water, such as ships.
- Rail transport. Vehicles which use rails to move.
- Road transport. Vehicles on the road.
- Specific. This concerns rumour about fire, a fire drill, fire alarm after a fire has been extinguished.

These concepts are referring to domains and not to different types of objects which can be on fire. For example, 'Aviation' can not be on fire, an 'aircraft' can. Because we wanted to create an ontology with homogeneous concepts and relations we changed these domain concepts.

The domain concept *Specific* in the emergency call categorization is used for calls about situations which don't need immediate actions. The concept *Specific* was made an *attribute* because the concepts subsumed under *Specific* are specifications of a relevant object. Figure 6.1 shows the backbone. The backbone consists of eight concepts referring to the domains which are subordinate concepts of *Object* as stated in STOR (see Chapter 2).

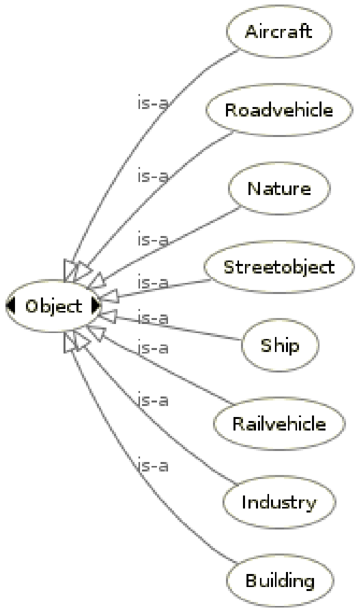


Figure 6.1: The backbone (domain concepts) for all the ontologies.

6.3.2 The Use of Frequency Data

To construct the expert-based and basic-level ontology we made use of frequency lists. We made use of corpora to derive the 'frequency of use' for specific words. Words which are frequently used, are words people are more familiar with than words not used that often. Because the ontologies we develop are used for generating questions to ordinary people, familiarity with words as represented by frequency of use, is of importance.

Corpus	**Entries**	**Nouns**
Subtlex	554.338	223.877
E-lex	615.120	244.896
Celex	380.581	166.619
Total unique	765.565	389.278

Table 6.2: Number of entries and nouns in used corpora.

We made use of three frequency lists as can be seen in Table 6.2. The most important for us was Subtlex which is based on a corpus of about 40 million words gathered from more than 8000 documents with subtitles (Keuleers et al., 2010). We also used e-lex (Van der Wouden et al., 2002) which was gathered from spoken text. The last list was from the Center for Lexical Information (Celex) (Baayen et al., 1995). This list contains frequencies based on a corpus of 42 million words taken from written texts (Baayen et al., 1993). We used the three lists to make our usage of frequency more robust .

Of all three lists we normalized the frequencies of lemmas on 10 million tokens. We encountered a number of problems using the frequency. Because the plural form used in AATNed (see Section 6.3.4) sometimes coincides with a verb, the frequency of the verb instead of the noun is extracted from the list. To prevent this problem we use the 'part of speech' which gives an indication of the syntactic role of a word. Furthermore, frequencies vary widely between the three lists, sometimes up to two orders of magnitude for the same concept. Because the frequencies showed such a great variance we took the mean of the three frequencies to work with.

6.3.3 Construction of an Ontology from a Pragmatic Categorization

The first ontology we developed was constructed from a vocabulary used by the fire department to determine the situation during an emergency call. When confronted with an emergency situation people are urged to call 112 (or 911 in the US). The vocabulary is used to interpret the message of the caller. From the vocabulary we constructed an ontology. This is what we call a pragmatic-based ontology. The knowledge in this vocabulary was gathered over a long period of time and is implicit. The classification in use is not directly suitable for automated reasoning. First, the relations are heterogeneous, for example, both *part of building* and *office building* are classified as subordinate concepts of *building*. Second, some of the terms used are not functionally the same, e.g. both *meeting* and *office building* are classified as *building*. Furthermore, situations are subordinate concepts of objects, e.g. *crash in water* and *crash on land* are both subordinate concept of *aviation* as are *number of passengers* in this classification.

We used a method described by van Assem (van Assem, 2010) to convert the vocabulary into an ontology in four steps. The first step is to analyze the vocabulary with the help of a conceptual model showing how concepts are related. This analysis made clear that most relations between the concepts in the pragmatic categorization are *is-a* relations. A few other relations were analyzed as being different from an *is-a* relation and perceived as *hasPart* and *hasNumber*.

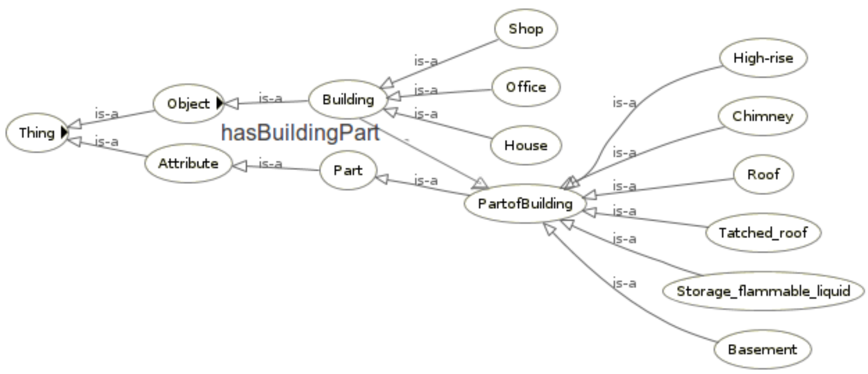


Figure 6.2: Part of the ontology constructed from the classification of the fire department. Shown are the *is-a* relations e.g. between *building* and a part of its subclasses and *PartOfBuilding* and all of its subclasses. Also the *hasBuildingPart* relation between *Building* and *PartOfBuilding* is shown.

The second step is the syntactic conversion from the source representation to a formal representation. Several guidelines for this conversion are mentioned by van Assem: use a basic set of RDF(S) constructs for the structure-preserving translation, preserve the original naming and avoid interpretation as much as possible.

The third step is a semantic conversion and interpretation when necessary. In this step we created a concept *attribute* subsuming some attributes of the relevant objects. These are *part of building*, *part of ship*, *number of passengers*, *location* and *location of crash*. The subordinate concepts of *attribute* kept a relation with the domain concepts (the backbone) to which they were also related in the original classification. These relations are: *has building part*, *has ship part*, *has location*, *has passengers* and *involved in crash*. Some names required changing because it was a description of a function instead of an object, for example we changed *meeting* to *conference center*. Also conjunctions and disjunctions were transformed into separate concepts. Furthermore, we added the concept 'train' as subordinate concept for the backbone concept 'rail vehicle', while otherwise this backbone would be without subordinate concepts. For the fourth step, which is standardization, we used OWL to standardize the representation of the classification we made.

A part of the resulting ontology is shown in Figure 6.2. Only three of the 13 concepts in the concept tree *building* are shown. All of these concepts have an *is-a* relation with *building*. Furthermore, *building* has a relation *hasBuildingPart* with subclasses of *part of building*.

6.3.4 Constructing an Expert-based Ontology

For the construction of the expert-based ontology we used two extensive sources for the Dutch language: the Dutch version of AAT (Art and Architecture Thesaurus) (AATNed, 2012) and Cornetto (Vossen et al., 2008) which is a Dutch WordNet-version. The Dutch version of AAT (AATNed) strongly resembles the original AAT. It has 34 hierarchies, i.e. homogeneous groups of concepts, 2,873 guide terms, i.e. superordinate abstract descriptions, and 30,817 concepts. Subsumption is formalized as skos:broader/narrower relationships (Bechhofer and Miles). Each concept has a preferred label which is in plural form when possible. Preferred labels which are homonyms are distinguished by qualifiers. An example from the English AAT is 'buckets (vessels)' and 'buckets (construction equipment)'. Many concepts in AATNed also have one or more alternative labels although much less than in the English AAT (24,817 vs 92,089) (Tordai, 2012).

Just like WordNet, Cornetto is a large database for language, not for English but for Dutch. It has no specific domain. Cornetto contains 70,370 synsets and 103,762 labels which is less than WordNet (Tordai, 2012). The hierarchy in Cornetto is much shallower than in AATNed.

The first step to construct an expert-based ontology based on AATNed and Cornetto was to map the backbone to several root concepts in both ontologies. Furthermore, we detected several 'orphans' in AATNed and Cornetto, i.e. concepts which have no superordinate concepts in any hierarchy. These orphans were also added to the initial set of candidate concepts for the resulting ontology. This resulted in a list of concepts from AATNed and a list of concepts from Cornetto as shown in Table 6.3 (AATNed/Cornetto).

The second step was to combine the concepts which were present in AATNed

Domain	AATNed	Cornetto	AATNed mapped	AATNed not mapped	Cornetto mapped	Cornetto not mapped	ebo
Building	2580	814	263	2317	284	530	3110
Industry	187	342	40	147	46	296	483
Nature	1280	623	78	1202	89	534	1814
Road vehicle	207	327	85	122	87	240	447
Aircraft	80	81	30	50	33	48	128
Ship	654	249	137	517	135	114	768
Rail vehicle	59	54	14	45	13	41	100
Street object	2410	3808	414	1996	494	3314	5724
Total	7457	6298	1061	6396	1181	5117	9574

Table 6.3: Number of concepts per domain in AATNed, Cornetto and the constructed expert-based ontology (ebo).

or in Cornetto but not in both the vocabularies (AATNed/Cornetto not mapped). These concepts and their superordinate concepts are an integral part of the new ontology. The third step was to give the concepts which are both present in AATNed and Cornetto a place in the ontology (AATNed/Cornetto mapped). In case of ambiguity we looked for similarity in their inheritance trees using two disambiguation techniques from Tordai: child and parent matching (Tordai, 2012). Child matching assumes that concepts which have a similar meaning also have a more or less similar hierarchical structure downwards. The concepts with the most child mappings are chosen. With parent matching we looked for mapping of the hierarchical structure upwards from the concepts with similar terms. Concepts that could be mapped unambiguously were merged and added to the ontology. Concepts which could not be mapped were also added to the resulting ontology.

As can be seen in Table 6.3 the ontologies are rather complementary: only 2182 concepts of the total number of concepts, which is 13755, were mapped, i.e. 15.86%. The large number of concepts in the domain *street object* is largely due to the subordinate concepts of the concept *persons* which has a great number of subordinate concepts such as *lawyer* and *walker*.

An example of the resulting expert-based ontology, i.e. a part of the concept tree *Building*, is shown in Figure 6.3. *Road transport buildings* is subordinate to *Transport buildings*, which is subordinate to *Transport constructions*. This sequence of subsumptions makes clear that the expert-based ontology has a large path length. It demonstrates the variety of abstraction at one hierarchical level. For example, *terminus* and *bus shelter* should be both subordinates of *bus station* while in this concept tree all three concepts are at the same level of abstraction. A related but slightly different problem is the imbalance of this ontology. It is unbalanced because sometimes one concept is superordinate to a concept tree, i.e. *toll constructions*, while others are not further deepened. Furthermore, concepts such as *mansiones* and *mutationes* will hardly be known

Figure 6.3: Part of the expert-based ontology.

by ordinary people (they are respectively roman rest houses and horse changing stations between them). These concepts are only in use by a small group of experts.

6.3.5 Restricting the Expert-based Ontology to Concepts with a High Frequency

To overcome the meager suitability for the task of generating questions and compliance with human categorization (see Section 6.5) we removed certain subtrees such as ‘person’ and ‘animal’ which resulted in the restricted expert-based ontology (‘rebo’). Furthermore, we enriched the concepts in the expert-based ontology with the frequency counts. One of the characteristics of basic-level concepts is their high frequency of use (see Section 6.2). When the low frequency words are left out of the expert-based ontology, the remaining concepts have their high frequency of use in common with basic-level concepts. We then formulated thresholds for each domain and used for each domain distinct minimum frequencies which resulted in the restricted expert-based ontology with thresholds (‘rebotd’). Here we encountered the problem that words can have more than one meaning, i.e. words are homonyms, and frequency lists do not distinguish between different meanings of the same word. For this problem of homonymity we had two strategies. The first strategy was to look for an English counterpart in the original AAT. When the English word had no homonyms we took the frequency of that word. When this strategy did not work we used a heuristic method and looked for the frequency of the siblings of the word. When the frequency of the word we were looking for was significantly greater than the frequency of its siblings this was considered to be an indication that we found another interpretation of the word than we were looking for. In that case we used the average frequency of the siblings. This resulted in the restricted expert-based ontology with thresholds and homonymy ('reboth'). As a last step the concepts with homonyms were left out which resulted in the restricted expert-based ontology with threshold and no homonymity ('reboth0').

6.3.6 Constructing an Ontology with Basic-level Concepts

As a third way to generate an ontology we elicited concepts and their attributes from ordinary people. For this we used the same method as Rosch et al. (Rosch et al., 1976a). First we asked participants to mention subordinate concepts of the eight backbone concepts mentioned in Section 6.3.1. Then we asked for attributes of these subordinate concepts. Furthermore, we present how we used these attributes to automatically generate a hierarchy based on basic-level concepts.

Experiment to gather basic level concepts and attributes

To detect subordinate concepts of the backbone concepts we asked 21 participants to name subordinate concepts. The participants were undergraduate students of the Amsterdam University of Applied Sciences. The questions the participants had to answer were:

1. Name different sorts of buildings. Example: school, office.
2. What sort of industry do you know? Example: petrochemical industry, food industry.
3. What sort of nature do you know? Example: heath, beach.
4. What sort of vehicles for on the road do you know? Example: car, lorry.
5. What sort of vehicles for in the air do you know? Example: plane, helicopter.
6. What sort of ships do you know? Example: ferry, freight ship.
7. What sort of vehicles for on rails do you know? Example: train, metro.
8. What sort of objects do you see when you walk/bicycle/drive on the road? Example: a building, sidewalk.

In Table 6.4 a summary of the results is shown. The number of distinct words mentioned by the participants is shown under C_m. First we normalized the mentioned words, i.e. we replaced plurals with singulars and diminutives with non-diminutives. Sometimes we had to correct some terms. In the list for industry a lot of language errors and inappropriate abbreviations occurred, e.g. auto was mentioned when probably auto industry was intended. We corrected these, i.e. *auto* became *auto industry*. The remaining words were enriched with their frequency of use found in the Subtlex-nl and e-lex list. Every word which was mentioned at least twice or had a frequency of more than one was taken

Domain	C_m	C_u	$\frac{C_u}{C_m}$	A_m	A_n	$\frac{A_m}{A_n}$
Building	53	47	0.89	4269	2487	1.72
Industry	52	19	0.37	1606	1012	1.59
Nature	32	20	0.63	1760	798	2.21
Road vehicle	41	19	0.46	1801	972	1.85
Aircraft	25	14	0.56	1234	654	1.89
Ship	48	24	0.50	1933	1090	1.77
Rail vehicle	39	15	0.39	1342	856	1.57
Street object	71	40	0.56	3709	1814	2.05
Total	361	198	-	17654	9683	-
Mean	45.23	24.78	0.55	2207	1201.04	1.86

Table 6.4: Results of the experiment to gather concepts and attributes. Concepts (C) which are mentioned (C_m) and used (C_u) and attributes (A) which are mentioned (A_m) or unique (A_n), gathered from participants.

as a subordinate concept. In the remaining list for the domain *road vehicle* we found the words 'trike' and 'quad'. These concepts were excluded from the set of subordinate concepts forming the domain *road vehicle* because they were not cited in the frequency lists. Of all the mentioned concepts we used 55% in our ontology. The number of concepts per domain which was used to gather attributes in the ontology can be found under C_u. The least subordinate concepts were mentioned for the domain *Aircraft*. Probably because participants are not familiar with this domain. The most familiar domains then appear to be *Building* and *Street object*.

In the second part of this experiment we asked 20 participants to mention the attributes of the concepts gathered by the experiment explained above. Participants were asked to name five most important visual attributes of an instance of the mentioned concept. In an information sheet the participants first were told the goal of the experiment, what basic-level concepts are and how they can differ per culture. Participants did the exercise at their own time mostly at home. All the attributes mentioned in association with a specific concept were collected in one set. The results are also shown in Table 6.4.

The number of attributes mentioned by the participants is shown as A_m. Many attributes were mentioned twice or more. The number of distinct attributes in each domain is shown as A_n. The mean number of times an attribute was mentioned is shown as $\frac{A_m}{A_n}$. The attributes of the subordinate concepts of *nature* are mentioned on average 2.21 times while the concepts subordinate to *rail vehicle* are mentioned on average only 1.57 times.

Finding a basic-level hierarchy

To detect superordinate concepts we used the concepts and their attributes as provided by the participants in the experiment described in the previous section. The result of this experiment is a list of all the attributes as mentioned by the participants for each concept. All 198 concepts and their attributes were used to construct a basic-level ontology by combining these concepts in such a way that the combination with the highest category utility (see Section 6.2.3) was detected. Each set of combined concepts was then used as the root concept of a new subtree for which the subordinate level was computed in the same way. This process stopped when all of the 198 concepts emerged at the lowest level of the hierarchy.

To create an ontology we used two algorithms: an algorithm which receives a set of domain concepts and their attributes as input (see Algorithm 7). This algorithm uses Algorithm 8 which returns the combination of concepts with the highest category utility. The original set of concepts is made root concept of this subtree. The concepts (which are a set of concepts) in the returned set are made subordinate concepts to the root concept. These subordinate concepts are input for Algorithm 8 to detect a new set of concepts with the highest category utility. These algorithms are analogous to the algorithms of Chen et al. (Chen et al., 2010) except for the computation of the category utility where we use Equation 6.1. All eight sets of domain concepts were processed by these algorithms. The results were added to the ontology as concept trees for the domains.

Algorithm 7 Ontology generation.

Input: set of concepts with their attributes (C_r)
if size of $C_r > 1$ **then**
  use C_r to get from algorithm 8 the set of subconcepts with the highest category utility(C_b)
  make C_r the root concept and every concept in C_b its subconcept
  use each subconcept in C_b as input for algorithm 7
end if

For Algorithm 8 we used a set of the concepts as input to detect the subset with the highest category utility. Category utility is a metric of the strength of the association between a concept and its attributes. Category utility (CU) of a concept (c) given a number of attributes (m) for that concept, $A = \{a_k\}_{k=1}^m$, is formalized as follows (Corter and Gluck, 1992):

$$CU(c, A) = P(c)\Sigma_{k=1}^m P(a_k|c)^2 - P(a_k)^2 \tag{6.1}$$

In this equation $P(c)$ varies, because $P(c)$ is the probability of occurrence of c in a determined set of concepts (C in Algorithm 8). This set of concepts gets smaller each iteration because the two most similar concepts are combined. $P(a_k)$ is the probability that in general a concept has attribute k. And $P(a_k|c)$ is the probability of attribute a_k given concept c.

Algorithm 8 Detecting superordinate concepts.

Input: Set of concepts with their attributes (C)
while size of $C > 1$ **do**
  step++
  compute similarity of concepts with eq. 6.2
  combine two most similar concepts into one concept (making one set of attributes)
  compute category utility with eq. 6.1
end while
Find the set of subconcepts (C_{CU-max}) with the highest category
Output: C_{CU-max}

Algorithm 8 iterates several times over the set of concepts and each time the category utility is computed and the two most similar concepts are combined.

For the computation of similarity we used the idf-cosine coefficient (Schultz and Liberman, 1999):

$$sim(a,b) = \frac{\sum_{k=1}^{n} idf(a_k) v_{a,k} v_{b,k}}{\sqrt{\sum_{k=1}^{n} v_{a,k}^2}\sqrt{\sum_{k=1}^{n} v_{b,k}^2}} \tag{6.2}$$

where $sim(a,b)$ stands for the value of similarity between the concepts a and b. The number of times a participant used this attribute to characterize concept a is indicated as $v_{a,k}$. $idf(a_k)$ is the inverse document frequency of the attribute k: $log(\frac{N}{N_{f_k}})$, where N is the total number of concepts and N_{a_k} the number of concepts having attribute k.

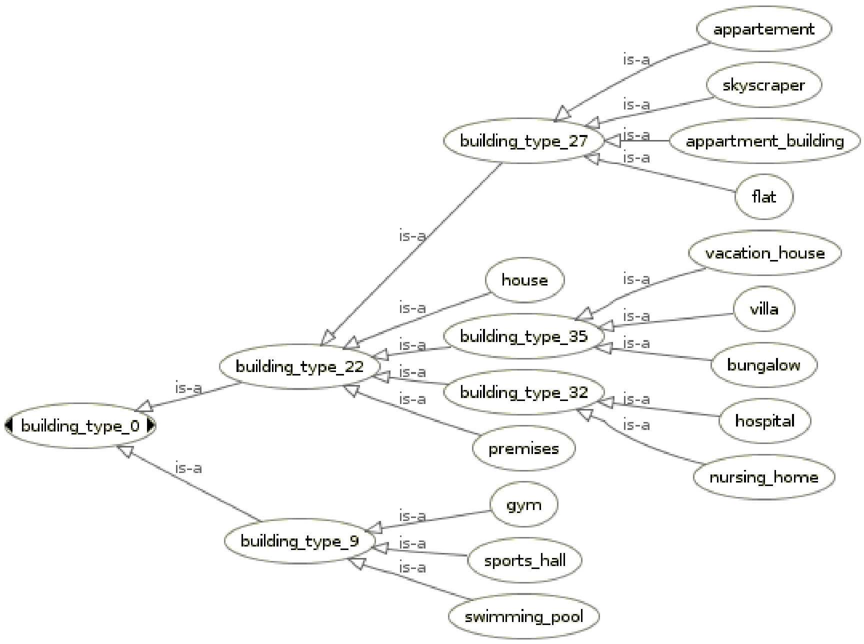


Figure 6.4: Part of the automated generation of the basic-level concept ontology. Superordinate concepts are named as their domain concept and suffixed with a number.

A part of the resulting ontology is shown in Figure 6.4. Only the domain concept *building* is shown and of this concept only 2 subordinate trees are shown. The superordinate concepts have no specific names and are called after their domain concept with a number as suffix to differentiate between them. Although the abstractions of the concepts have no meaningful name, they have distinguishing attributes. For *building9*, which is the concept for buildings where people are involved in sport, these attributes are *shower* and *dressing room*. For the concept *building22*, buildings people live or sleep in, the distinguishing attributes are *roof* and *room*. Concept *building27* are buildings with multiple floors. The distinguishing attributes are *floor* and *high*. For concept *building35*, which is the concept building associated with holidays or luxury, the distinguishing attribute is *swimming pool*. And *building32* is the concept for buildings especially for sick and/or old people with the attributes *counter* and *pharmacy*.

6.4 Evaluation Framework

To establish which ontology is most suitable for automatic generation of questions to determine a crisis situation, an evaluation framework is used. The framework consists of metrics which are used to reflect the characteristics of the ontologies and compare the results to the criteria we presented in Section 6.1. The construction of the framework is presented next. The results of the evaluation are presented in Section 6.5.

6.4.1 Structure

General metrics such as the number of concepts, maximum and average path length give a first impression of the ontology (Zhe et al., 2006; Lozano-Tello and Gómez-Pérez, 2004; Bolotnikova et al., 2011). For us they also signify the suitability for the task of generating questions in several ways. The number of concepts indicates how much information an ontology offers. The maximum and average path length of an ontology are directly related to the number of questions which can be generated from an ontology because with each subtree we generate a new question. The maximum and average number of subclasses are directly related to the number of answers offered to choose from with each question. The degree of coherence tells us something about the possibility of getting more questions than average. Coherence is determined by computing the number of paths which are extraordinarily long, i.e. the ratio of the number of paths which are longer than average plus two times the standard deviation to the total number of paths.

6.4.2 Completeness

The evaluate whether the concepts used in an ontology are appropriate one compares the conformity between the ontology and the domain of knowledge which the ontology tries to capture. The standard method in this approach is to use a 'gold standard' (Brewster et al., 2004). Since no reference corpus for the domain exists, we had to develop a gold standard. To determine whether the concepts people use to describe fire situations are part of our ontologies (coverage) and whether our ontologies contain concepts by people do not use (precision) we compare the ontologies with the reference corpus (Guarino, 2004).

To create a corpus we asked participants to write down observation statements while watching several videos about objects on fire. An observation statement is a statement which describes the immediate perception of a person, for example 'I see a building on fire'. From such statements the concepts 'building' and 'fire' are extracted by selecting the nouns from the observation statements. We collected 8 series of videos about subjects related to the domain as determined in Section 6.3.1. These videos were found on the Internet, most of them were taken by amateurs and shown in their original form. The duration varied from one minute to a quarter of an hour. The instructions given to participants were[2]:

[2] All of the experiments mentioned in this article were done in the Dutch language which is the native language of most of the participants.

> You are asked to write down observation statements. Observation statements are sentences like 'I see a car hitting another car', 'it is night', 'I see there are lanterns'. When you are watching a video and you want to write an observation statement, stop the video and write the statement.

To collect observation statements 21 students from the Amsterdam University of Applied Sciences wrote down observation statements while observing all videos in several separate sessions. The participants took their assignments home and did their task in their own time.

Domain	**Nr. of videos**	**Nr. of words**	**Nr. of different words per domain**	**Nr. of different nouns per domain**	**Unique nouns per domain**	**Perc. unique nouns per domain**
Building	9	21095	1589	267	38	14.23
Industry	8	15128	1164	245	26	10.61
Nature	9	20290	1404	285	40	14.04
Road vehicle	8	20762	1551	277	37	13.36
Aircraft	9	16733	1244	241	47	19.50
Ship	8	13713	1043	200	29	14.50
Rail vehicle	9	19505	1462	346	61	17.63
Street object	9	9540	949	217	19	8.76
Total	69	136766	10406	2078	297	14.33

Table 6.5: Results of the experiment to elicit observation statements and generating a corpus.

The results of this experiment are shown in Table 6.5. With eight series of videos we collected 136,766 words. Not surprisingly the most used words were 'fire' (2,056) and 'smoke' (1,596). Words in plural form were substituted by their singular form and diminutives were replaced by their normal form. This resulted in 10,406 distinct words for all series of videos. For example, in the series of videos about buildings on fire 1,589 different words were used by our participants. On average only 14.33% of the nouns used in a series of videos are unique to that series of videos.

It is clear that an ontology with a high coverage and a high precision when compared with the reference corpus is desirable. Precision is often less important and thus low precision and high coverage is less desirable. When the coverage is low and precision high, this is not good but a low coverage and low precision is even worse (Guarino, 2004). The comparison of the ontologies with the gold standard can be found in Section 6.5.3

6.4.3 Cognitive Semantics

Whether the formalized representation of knowledge complies to human categorization, is presented by three metrics as shown in Table 6.1 in the column 'Cognitive Semantics'. A combination of path length and number of subordinate concepts is represented by entropy which is described shortly. This gives a more refined indication of the amount of information in an ontology than the number of concepts. Entropy as we interprete this, gives an indication of the number of questions one has to answer and the number of answers one has to choose from, i.e. how much information one has to process to determinate a situation. How much information a human can process at one time is reflected by the metric 'cognitive ergonomics' for which we used the Ingve-Miller number. The last metric we developed to characterize the meaning of an ontology is 'semantic distance validation'. This metric signifies how the categorization offered by an ontology complies with the way people categorize concepts. The experiment to determine the differences between the three ontologies in this respect concludes this section.

Entropy

We gather information by asking questions. Each answer increases the amount of information about the situation we are interested in. When we ask for what kind of object is on fire, a more specific answer such as 'toll houses' is giving more information than a general answer such as 'road transportation buildings' (see Figure 6.3). By successively asking questions, generated from an ontology, the amount of information is increasing. The number of questions we can generate from an ontology is an indication of the amount of information our system using that ontology can elicit from a user. We are using this metric to compare the ontologies we created.

A determination of the amount of information is entropy which is defined as the information content of a message or the uncertainty a receiver has about the message from a sender before receiving it (Shannon, 1948). This is calculated as:

$$H(X) = -\sum_{i}^{n} p(x_i) log\ p(x_i) \tag{6.3}$$

where $p(x_i) = Pr(X = x_i)$ and X is a discrete random variable.

The entropy of an ontology can be computed by taking the number of relations a concept has and sum over the probabilities of these relations (Calmet and Daemi, 2004; Cho et al., 2007; Doran et al., 2009; Resnik, 1995, 1999). It is a metric based on the information content of an ontology (Scharrenbach et al., 2010; Coskun et al., 2011; Palmisano et al., 2009). Cho et al. take entropy to be the ambiguity caused by the number of subsumption relations a concept has (Cho et al., 2007). The informational content of a concept i is determined as the ratio of the number of all members of a concepts subtree (including itself) $|\Phi_i|$ to the total set of concepts $|\Omega|$. When $\Phi_i \subset \Omega$,

$$p_i = \frac{|\Phi_i|}{|\Omega|} \tag{6.4}$$

The information content of a concept in an ontology is then used as $p(x_i)$ in Equation 6.3. The amount of information increases proportionally with the number of nodes in the ontology.

Using an ontology with a higher entropy than the entropy of another ontology means that our system can generate more information under the assumption that the user knows about all the differences between the presented concepts. Ordinary people can only to a certain degree give information about a situation because they are limited to a specific level of detail due to their knowledge of the objects involved in the situation. Domain experts on the other hand, are capable of giving much more information about the same situation when using the same ontology.

Each of the three ontologies we developed, has a specific entropic value related to the method of creation. Because the level of detail determines the level of entropy, we expect the entropy to be high for an ontology based on expert knowledge and low for the pragmatic-based and basic-level ontology. We use the ontologies for the task of generating questions for ordinary people, thus it is not necessarily so that the ontology with the highest entropy is the best ontology.

Ingve-Miller number

It is well known from psychology that humans have a limited channel capacity, i.e. the extent to which the observer can match his responses to the stimuli given to him. Experiments show that no matter how much different signals a person gets he can only differentiate between seven plus-minus two concepts: the Ingve-Miller number (Miller, 1956; Bolotnikova et al., 2011). Although we do not present an ontology to humans directly, the structure of an ontology will be reflected in the number of possible answers generated from an ontology. The number of answers has influence on the time needed to choose the right answer, i.e. when someone has to choose the right answer from a list of 15 concepts it will take longer than when someone has to choose from a list of 5 concepts.

The Ingve-Miller metric reflects how comprehensible an ontology is for a human and computes the number of concepts directly subordinate to that concept. The Ingve-Miller metric for breadth is the ratio of number of concepts i having a number of directly subordinate concepts ($|\Psi_i|$) which does not exceed the Ingve-Miller number to the total number of concepts: $IM_b = \frac{\sum_{i=1}^{m} |\Psi_i| \leq 9}{|\Omega|}$. The closer this ratio is to 1, the higher the cognitive ergonomics.

Semantic distance validation

To evaluate how the structure of the ontologies complies with how people categorize their knowledge we designed an experiment. In this experiment we determined the match between the formal semantic distance of concepts within an ontology and the cognitive semantic distance of the same concepts as perceived by the participants. This experiment is analogous to psychological experiments such as done by Smith and Kemlar Nelson (Smith and Kemler, 1984) and the knowledge eliciting technique known as repertory grid (Kelly, 1955; Gaines and Shaw, 1993; Schreiber et al., 1999).

In an experiment participants are shown sets of three concepts, i.e. triads, generated from one ontology. Of these three concepts two are subsumed by the same superordinate concept at some level of the ontology and one is not, i.e. the 'outsider concept'. The largest distance between two of the three concepts subtracted by the smallest distance between two of the three concepts is what we call the semantic distance. We generate randomly numerous triads with various semantic distances. Participants are asked to determine which concept is the outsider concept. We expect the participants to have more difficulty with triads which have a small semantic distance than triads which have a large semantic distance. The instructions to the participants were:

> In the files which are attached you find sets of three concepts. Of these three concepts there are always two which are more similar than the third concept. Choose the concept which does not belong to the other two.
> An example:
> a. forest
> b. meadow
> c. freighter
> # 34:
> You now choose from these three concepts. It seems clear that in this case you choose 'c. freighter'. Thus you write # 34: c.

This experiment was done by participants from the Amsterdam University of Applied Sciences. The participants took their assignments home and did their task in their own time. The results per ontology are categorized according to the semantic distance and are presented in Section 6.5.4. When compared with each other, these results will give an indication of the difference between the formal and cognitive semantic distance of concepts in the respective ontologies. A high number for an ontology indicates a better compliance with human categorization than a low number.

6.5 Evaluation of the Ontologies

The results of the metrics are used to evaluate the four criteria we have formulated. Each criterion is discussed in a separate section. The results are discussed in Section 6.6 and a conclusion is drawn in Section 6.7.

6.5.1 Structure

Characteristics of the ontologies have been determined and are summarized in table 6.6. The pragmatic-based ontology (pbo) is the smallest ontology with 93 concepts. The expert-based ontology (ebo) is the largest ontology with 9574 concepts. It is obvious that restricting the ebo decreases the number of concepts. When, for example, all roles people can have such as 'student', 'lawyer', 'friend', and all different animals are left out the number of concepts remaining is about one half of the original ebo (see Section 6.3.5. The basic-level ontology (blo) keeps the middle with 243 concepts.

The pbo also has the smallest maximum (3) and average path length (2.05) which means this ontology is very shallow. The ebo has the greatest maximum and average path length. The average path length of the ebo is four times longer than the pbo which means on average four times more questions are needed. Restriction of the ebo by only using concepts with a frequency above a threshold

Metrics	ebo	rebo	rebotd	reboth	reboth0	pbo	blo
Nr of concepts	9574	5409	1606	1045	773	93	243
Max. path length	16	16	15	15	15	3	4
Av. path length	9.29	9.08	8.74	8.70	8.61	2.05	3
Coherence	0.0186	0.0148	0.0177	0.0166	0.0143	0	0
Max. nr. of subcl.	446	86	49	47	30	19	11
Av. nr. of subcl.	4.67	4.40	2.65	2.57	2.19	7.67	3.55
Coverage	0.46	0.37	0.35	0.30	0.26	0.12	0.22
Precision	0.04	0.05	0.17	0.23	0.27	1	1
Entropy	35.57	33.98	29.11	27.56	26.82	6.42	10.67
IM breadth	0.902	0.902	0.967	0.969	0.976	0.75	0.985

Table 6.6: Summary of the metrics for the expert-based ontology (ebo), restricted expert-based ontology (rebo), restricted expert-based ontology with threshold (rebotd), restricted expert-based ontology with threshold and homonymy (rsbthf), restricted expert-based ontology with threshold and without homonymy (reboth0), pragmatic ontology (pragm) and basic-level ontology (basic).

and eliminating homonyms improves the ebo. The blo has a maximum path length of 4 and its average path length is 3 which makes it a shallow ontology.

The coherence for the pbo and blo is 0 which means these ontologies do not have extraordinarily long paths. All the expert-based ontologies do have such outliers which means that it can happen that using one of these ontologies generates substantially more questions than the average. The number for coherence indicates the possibility one has to determine an object, when using our application which generates questions from an ontology, while the path to the representing concept is extraordinarily long. For example, when using the reboth0 there are (1.2% of 773 $\approx$) 9 concepts which have a path substantially longer than the average path.

The maximum number of subclasses for the ebo is 446 which makes it less suitable than other ontologies for the task of question generation. A restriction of the ebo reduces the maximum number of subclasses but it never becomes as small as the pbo or blo. For the average number of subclasses the expert-based ontologies show better values. For the pbo the average number of subclasses is 7.67, which is the highest value compared to the other ontologies. This high value for the average number of subclasses is a result of its shallowness.

The shallowness of the pbo and blo is an advantage. This advantage is (partly) reduced by the larger average number of subclasses than for the restricted expert-based ontology. Moreover, the average number of subclasses is influenced for the better by the long paths of expert-based ontologies which represent many superordinates with one subordinate.

6.5.2 Efficiency of Construction

The efficiency of construction varied among the three ontologies. The construction of the basic-level ontology costs the most time of all three constructions we carried out. To elicit subordinate concepts and attributes from participants is a time consuming endeavour for which an experiment had to be done. The automatic construction of a hierarchical structure based on the attributes is a relatively small effort knowing the algorithm. A disadvantage of the way we constructed the basic-level ontology is that the superordinate concepts have no names.

When a pragmatic categorization of a particular domain is available the conversion to a pbo is efficient but still costs some time. The expert-based ontology was easy to construct once the method was clear because it was done in an automated way. When knowing how to do this the efficiency of constructing such an ontology is the highest of the three ontologies.

6.5.3 Completeness

A comparison between the ontologies and our corpus did not show a clear difference between the ontologies. We then extended the corpus with the concepts used in the pbo and blo. All these concepts have in common that they are elicited directly from ordinary people. Because the superordinate concepts in the blo are automatically generated we did not use these. The results are shown in Table 6.6.

The ebo is the most encompassing ontology, i.e. the ontology which has the most concepts also found in the gold standard. The coverage is 46%. The precision of the ebo is not high: only 4% of its concepts can be found in the gold standard. Restricting the ebo reduces the coverage but it remains above the coverage of the pbo and blo. The restriction of the ebo enhances the precision. The precision of the pbo and blo is due to their incorporation into the corpus 100 % (see Section 6.4.2), their coverage is respectively 12% and 22% which is a lower number than the coverage of the ebo.

6.5.4 Cognitive Semantics

The entropy of the pbo is the smallest due to its shallowness and small number of concepts. This means the information content of the pbo is smaller than the information content of other ontologies and there is a strong distinction between the concepts. The entropy of the ebo is the highest of the three ontologies. Restricting the ebo in various ways reduces its entropy but it remains considerably higher than the entropy of the pbo and blo. The entropy of the blo is smaller than the rebo but higher than the pbo.

The results of the Ingve-Miller number of the pbo are poor: the Ingve-Miller number for breadth is the lowest (0.75) compared to the other ontologies (0.902 and 0.985). Restricting the ebo increases this number.

The semantic distance validation of the pbo shows a mean of 70.84% which is low compared to the blo which scored 79.83% (see Table 6.7). When we look at the triads with a distance greater than 2 the mean gets better (82.71%). The semantic distance validation for the ebo is only 51.42%. Restricting the ebo increases the number for the semantic distance validation. The validation of the hierarchical structure of the concepts used in the 'restricted expert-based with threshold and no homonymy' shows that this ontology complies better with human categorization than the original ebo.

The validation of the hierarchical structure clearly shows that the blo has a categorization which complies best with how ordinary people categorize. It also shows that the pbo has a disadvantage because of its small maximum path length. The cognitive semantics of the blo is best on the Ingve-Miller number and semantic distance validation. When comparing the ontologies on their en-

Semantic distance	pbo (N = 73)	ebo (N = 57)	reboth0 (N = 48)	blo (N = 73)
1	35.24	51.67	45.48	42.35
2	69.28	59.76	61.37	82.33
3	83.10	61.87	54.60	81.88
4	95.75	47.71	63.97	88.30
5		54,40	64.23	85.35
6		50.78	49.05	98.76
7		64,78	57.93	
8		64,65	86.49	
9		7.14	70.32	
Mean	70.84	51.42	61.49	79.83

Table 6.7: Percentage right answers from the experiment to validate the categorizations (see 6.4.3). Semantic distance indicates the minimum number of steps between the 'outsider concept' and the concepts which belong to the same superordinate concept.

tropy it can be seen that using the pbo elicits the smallest amount of information and the (restricted) ebo the largest amount of information.

6.6 Discussion

Many domains are more or less successfully captured in pragmatic classifications or expert-based ontologies. But these ontologies do not fulfil the criteria we have formulated. Most ontologies are not constructed with the goal to comply with human categorization. Specific biases are another problem with these existing ontologies. These biases make them unbalanced and their structure incoherent. For example, the backbone concept *ship* not only refers to all sort of ships but is also superordinate to concepts such as *swimmer*, *pipeline* and *offshore platform*. Another example is the bias of the expert-based ontology towards a few domains, i.e. architecture and ships, which is caused by the historical roots of AATNed. The construction of AATNed is in part based on a large vocabulary of maritime concepts.

To construct the expert-based and basic-level ontology we used frequency counts of words from Subtlex, e-lex and CELEX. Because of the homonymity of words, such word lists are often not reliable. In a frequency list it is often impossible to find the frequency for the word with the intended meaning.

A major drawback of the basic-level ontology is its lack of meaningful terms for the superordinate concepts. Replacing unnamed superordinate concepts in the basic-level ontology by superordinate concepts from the expert-based ontology could compensate this drawback of the basic-level ontology in some cases.

The creation of the gold standard is open for improvement. A major part of it consists of concepts elicited from people in an experiment during which they made observation statements of fire situations. The participants of the experiment used a lot of words we did not find in the ontologies. The explanation for this discrepancy is that we gathered words for the corpus which are probably related to another task. Instead of focusing on the object of interest, people tend to describe more of the context. When people are asked to give information about a situation as in our experiment, they are not challenged to give detailed descriptions of what they see. This supports the suggestion that to gather information about a situation people have to be asked specific questions to determine that situation to get a clear and detailed understanding of what is going on.

One of our metrics to find an ontology which complies with human categorization is entropy. Entropy is an indication of the information content an ontology offers. In contrast to the Inge-Miller number, which is well documented in psychological literature, it is to date not possible to state the level of entropy suitable for ordinary people. There are some indications though. Cho et al. (Cho et al., 2007) found an average entropy of 17.48 for domain ontologies in WordNet. We found for the basic-level ontology an entropy of 10.67 and for the restricted expert-based ontology without homonyms an entropy of 26.82. We would like to suggest that the optimal level of entropy for ontologies suitable for ordinary people lies somewhere between those numbers.

Furthermore, we expect that the evaluation framework as presented here can be applied to other tasks. The criteria we stated are independent from the metrics we used and will be different when using ontologies for another task. The metrics we used for an indication of the cognitive semantics can be used to evaluate ontologies regarding all tasks which demand a compliance of the ontology with human categorization.

6.7 Conclusions

To elicit information from ordinary people we automatically generate questions about a situation. These questions are derived using an ontology. Because the quality of the questions is influenced by the ontology it is important to create an ontology which is compliant with human categorization. We first established a backbone of domains, then created three different types of ontologies (a pragmatic-based, expert-based and basic-level ontology) and formulated metrics for structure, completeness and compliance with human categorization. Furthermore, we evaluated the efficiency of construction. A summary of this evaluation is shown in Table 6.8.

We constructed the pragmatic-based ontology from a classification in use by the fire department. We adjusted the categorization by hand following a specified method which is an efficient way to construct an ontology. This ontology scores very low on completeness and the validation of the semantic distance. So, the pragmatic-based ontology only satisfies two of the four criteria: structure and efficiency of construction.

The expert-based ontology which we merged from AATNed and Cornetto is the most complete ontology but its compliance with human categorization is the worst relative to other ontologies. To overcome this problem we restricted this ontology to concepts which are frequently used. The construction with merging techniques is done relatively quickly. The expert-based ontology satisfies two of the four criteria: efficiency of construction and completeness.

To construct the basic-level ontology we elicited concepts from ordinary people. With these concepts and their attributes an hierarchy of concepts was automatically generated. The basic-level ontology scores best on the criterion compliance with human categorization but is the most expensive to construct. The high score on semantic distance validation for the basic-level ontology suggests that the algorithms used to create the basic-level ontology are suitable for this purpose. The basic-level ontology satisfies two of the four criteria: structure and cognitive semantics.

	pbo	**reboth0**	**blo**
Structure	+	-	+
Efficiency of construction	+	++	- -
Completeness	- -	+	-
Cognitive semantics	-	- -	++
Overall	-	+/-	+/-

Table 6.8: Summary of evaluation per criterion of the pragmatic-based ontology (pbo), restricted expert-based ontology with threshold and no homonymity (reboth0) and basic-level ontology (blo). Overall gives an indication of the overall evaluation.

We compared the ontologies with each other on one particular criterion and evaluated the results accordingly (see Table 6.8). For the criterion regarding the structure of the ontologies we evaluated Table 6.6. All of the metrics except 'average number of subclasses' show relative high numbers for the reboth0. This resulted in a '-' for the reboth0. For the efficiency of construction we evaluated the pbo and reboth0 as good and the blo as bad because of all the required experiments with participants which is very time consuming. For the construction of the pbo a vocabulary had to be re-engineered which is more work than just running an algorithm as can be done to construct reboth0. Reboth0 is the most and pbo the least complete ontology as can be seen in Table 6.6. The best ontology with respect to cognitive semantics is the blo: it has the highest number for the Ingve-Miller number and also for the semantic distance validation. Its entropy is not the smallest but still very low. The reboth0 has the lowest value for semantic distance validation and its entropy is the highest.

We see a continuum between the pragmatic-based and the expert-based ontology along the dimensions of length of dialogue to determine the situation and the amount of information gathered. With the pragmatic-based ontology the number of questions asked will be the smallest before a situation is determined, but the amount of information reflected by the entropy of the ontology also is the smallest. With the expert-based ontology the number of questions will be the largest before the situation is determined and the amount of information is the largest. The basic-level ontology and the restricted expert-based ontology are somewhere in between on this continuum. On this continuum one will have to look for an optimum determined by the specific task one sets and the amount of information one wants to gather. Entropy gives an indication for the choice to make on this continuum. When needing information in great detail and the user is capable of answering very detailed questions an expert-based ontology is preferable. When it is paramount to gather information quickly from laymen one should perform that task with a basic-level ontology.

Given the lack of sufficient data on basic-level ontologies in most domains we may have to rely on technical means to select basic-level concepts in expert-based ontologies. The coverage of the restricted expert-based ontologies shows that these ontologies can replace the basic-level ontology when constructing a basic-level ontology is not feasible. The coverage can further be enhanced by an expert editing the expert-based ontology to include missing concepts. An enhancement of the precision of expert-based ontologies can be achieved by manually excluding concepts which are not well known to ordinary people.

A combination of the expert-based and basic-level ontology might be an idea to work on when generating questions to ask about a situation. In our question generating application first an open question can be asked. The answer may refer to a concept at a certain level of detail in the expert-based ontology. This concept will then be the starting point of a closed question session generated from the basic-level ontology.

In summary, we have investigated three methods for creating ontologies to support question generation. Each of these methods has its merits, but from a practical point of view, in particular considering the required effort in constructing the ontology, using existing ontologies and subsequent editing of them seems to be the most promising.

Chapter 7

Experiments to Validate the Ontologies for Question Generation

In Chapter 6 three distinct ontologies were presented. It was shown, by evaluating the ontologies in a framework, that the resulting ontologies had specific characteristics which were related to the method of creation. In this chapter the ontologies are tested using the Situational Awareness Question Generator (SAQG), which was presented in Chapter 3, to generate questions for ordinary people who are part of a crisis situation. The questions are generated using the fifth strategy because the concepts represented in the ontologies are related by subsumption. The ontologies created in Chapter 6 are not immediately usable for SAQG. First some knowledge engineering has to be done. Two experiments have been performed to validate the pragmatic-based, expert-based and basic-level ontology. The results show that when a fast determination of a situation is required and the cost of development is not a problem the basic-level ontology is to be used and when accuracy is crucial the expert-based ontology should be used.

A shortened version of this chapter with emphasis on the second experiment was presented at BNAIC2014 (Teitsma et al., 2014b).

7.1 Introduction

In this chapter we present experiments with participants who determine what kind of object is on fire. With these experiments we try to identify the most suitable ontology to determine a situation and validate the metrics used in Chapter 6. When these metrics are valid the characteristics of the ontologies presented in the previous chapter also can be ascribed to the ontologies based on the results of these experiments.

The three ontologies we use in the experiments are created with three different methods (see Chapter 6). The pragmatic-based ontology (pbo) is developed from a categorization used at an emergency call center. The expert-based ontology (ebo) is constructed from existing ontologies created by domain experts. The basic-level ontology (blo) is an ontology based on empirical data elicited from ordinary people.

To test these ontologies we conducted two experiments. First a pilot experiment was done with a small group of participants (n=20). An evaluation of this experiment led to some methodological modifications in the second experiment. This experiment was done with a larger group of participants (n=110). The results of both experiments are compatible with each other, although the data of the second experiment gave rise to a more clear-cut conclusion.

The process of knowledge engineering to adapt the ebo and blo is presented in Sections 7.2.1 and 7.2.2. An evaluation of the resulting ontologies using the metrics presented in Chapter 6 is done in Section 7.2.3. The two experiments we conducted and their results are presented in Sections 7.3 and 7.4. The results are discussed in Section 7.5 and conclusions are drawn in Section 7.6.

7.2 Re-engineering Ontologies

To make the ontologies suitable for the application two of the three ontologies we created in Chapter 6 had to be re-engineered. The ebo could easily be made more efficient for use in SAQG. The blo was not suitable because most superordinate concepts did not have a meaningful label. Only the pragmatic-based ontology could be used by the application without modifications.

7.2.1 Expert-Based Ontology

We used the restricted expert-based ontology with threshold and zero homonyms (rsboth0) as this was the ontology most suited for our application. The ebo as created in Chapter 6 was significantly improved by applying some small changes. The most important of these improvements was the reduction of path length for some concepts. This reduction was applied when a superordinate concept only had one subordinate concept. The superordinate concept was then replaced by the subordinate concept. The way we use an ontology in our application, i.e. asking a question and suggesting answers, does not generate information when only one answer is possible. When a concept was subordinate to a concept with more or less the same label such as 'religious building' and 'religious building structure' the superordinate was replaced by that subordinate concept. Some superordinate concepts were split in two superordinate concepts because there was a clear difference between one group of concepts and another. We joined some synonyms and some concepts were replaced as subordinate to another concept than in the original ontology when this seemed appropriate. An example of this last adaptation was 'Navy ships' which was subordinate to 'Ships by function'. This concept was made subordinate to 'Vessel' and 'Ships by function' was removed. Figure 7.1 shows the partial result for the subtree 'Water vehicles'.

7.2.2 Basic-Level Ontology

The most important deficiency of the blo is the lack of labels for the superordinate concepts. Using the attributes of concepts to make a hierarchy of concepts results in superordinate concepts without a meaningful label. The concepts were labelled with the name of the domain and an enumeration suffix, for example

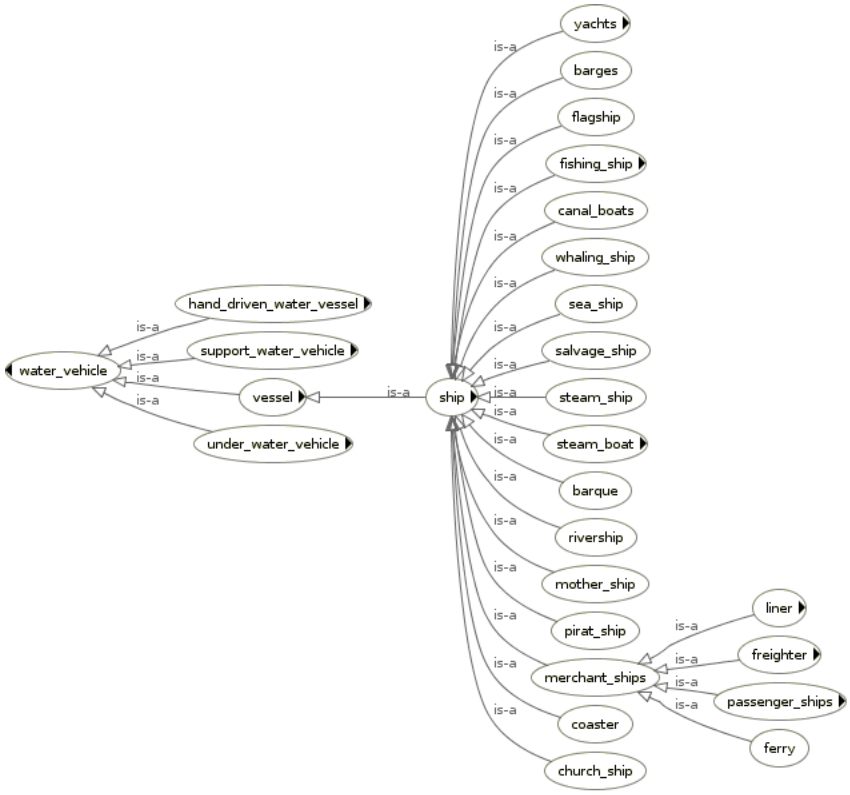


Figure 7.1: Subtree Water vehicles of the new expert-based ontology.

'building15'. Especially for the application of this ontology in SAQG an absence of meaningful labels is unacceptable.

The superordinate concepts without a meaningful label had to be given a label which referenced all the subordinate concepts. Such labels preferable are drawn from natural language and should be a meaningful generalization of the subordinate concepts. Labels fulfilling these two requirements are sometimes hard to find which made it necessary for us to create labels less natural than we would like. To create labels for the superordinate concepts we made use of guidelines formulated by van Heijst (van Heijst, 1995). These guidelines are:

1. decide whether the concept is sufficiently general to cover the pieces of knowledge that will be modelled using the concept,

2. decide whether the concept is sufficiently specific to only cover the pieces of knowledge that will be modelled using the concept and

3. decide whether the name of the concept is a meaningful term in the application.

Van Heijst uses these guidelines to find accurate concepts in a library of concepts. Our library is the natural language, so to speak. From this library we draw a concept which has the same meaning as the superordinate concept with the meaningless label. The meaning of the concept with the meaningless label is a generalization of the subordinate concepts. It is clear that the most perfect label consists of one word. We used several ways to give superordinate concepts a label:

- using the attributes. For example: ‘Holiday accommodation’ for the subordinate concepts ‘Bungalow’, ’Villa’, ’Holiday home’ because all these concepts had attributes referring to leisure activities such as ‘sun’, ‘holiday’, ‘swimming pool’, ‘rest’, ‘beach’.
- using labels from the ebo. For example: ‘Emergency service vehicles’.
- moving a subordinate concept one level higher up in the hierarchy. For example: ‘Office building’ which was subordinate together with ‘Office’ and ‘Government building’ to a superordinate concept without a meaningful label.
- using a shared function of the subclasses. For example: ‘Educational institution’ for the subordinate concepts ‘Library’ and ‘School’.

Unfortunately it was hard to find appropriate words for some of the superordinate concepts and we had to be content with compound labels for concepts such as ‘building where you can buy things’.

7.2.3 The New Ontologies

The ontologies created were given new names by adding the prefix ‘new’: new expert-based ontology (nebo) and new basic-level ontology (nblo). These revised ontologies were evaluated using the metrics presented in Chapter 6. The results of this evaluation are shown in Table 7.1. Several characteristics of the ebo changed for the better in the nebo. The number of concepts is reduced from 773 to 632. The maximum path length is reduced from 17 to 8. This also reduced the average path length from 8.61 to 3.60. The maximum number of subclasses is increased from 29 to 32, as is the average number of subclasses. This increase is a consequence of the reduction of the number of superordinate concepts with only one subordinate concept. Furthermore, the entropy of the expert-based ontology is reduced from 27.47 to 14.10. The nblo has only slightly different characteristics than the blo as created in Chapter 6.

	pbo	**nebo**	**nblo**
Number of concepts	93	632	226
Maximum path length	3	8	4
Average path length	2.03	3.60	2.71
Maximum number of subclasses	19	32	12
Average number of subclasses	7.15	3.85	3.59
IngveMiller number breadth	0.77	0.91	0.98
Entropy	6.43	14.10	10.43

Table 7.1: Metrics of the three ontologies used in the experiment.

To validate the semantic distance, i.e. the difference between the formal semantic distance between concepts within an ontology and the cognitive semantic distance as perceived by humans, we did the same experiment as described in Chapter 6 again for the newly created ontologies with 48 participants. The results can be seen in Table 7.2. For completeness the result for the pragmatic-based ontology as presented in Section 6.5.4 is added. The improvement of the score shows that the re-engineering was beneficial for the expert-based ontology. The re-engineering of the basic-level ontology did not lead to improvement on this metric. The reduction of the values for the nblo is due to the increase of concepts and the labels we added. Using the metric to validate the semantic distance, as described in Section 6.4.3, the superordinate concepts in the basic-level ontology were left out because they were without meaning. Another possible explanation is that not all the labels we added to these superordinate concepts were understood well.

Distance	**pbo** N=73	#	**nebo** N=48	#	**nblo** N=48	#
1	35.53	70	46.20	27	51.29	33
2	70.49	114	54.52	54	63.16	79
3	84.28	50	55.63	56	59.00	83
4	94.55	5	67.47	43	77.94	34
5			62.24	32	88.44	7
6			74.61	16	97.62	2
7			61.61	6		
8			90.63	3		
9			48.39	1		
Mean	70.84		63.54		72.67	
Mean(chapter 6)	70.84		61.49(ebor)		79.83	

Table 7.2: Results from experiment with triads.

In the experiment as described in the next sections we present participants with situations in three domains: ‘Buildings’, ‘Road vehicles’ and ‘Water vehicles’. In Table 7.3 some characteristics of these subtrees in each ontology are shown.

		pbo	**nebo**	**nblo**
Buildings	Number of concepts	15	198	56
	Average path length	1.71	3.68	2.72
	Average number of subclasses	5.33	3.26	3.56
	Entropy	2.67	9.53	5.74
Water vehicles	Number of concepts	20	127	21
	Average path length	1.77	4.08	3.04
	Average number of subclasses	7.00	4.40	2.75
	Entropy	2.94	7.50	3.92
Road vehicles	Number of concepts	15	97	22
	Average path length	1.71	3.59	2.88
	Average number of subclasses	5.33	4.59	3.83
	Entropy	2.67	7.02	3.80

Table 7.3: Metrics of the subtrees.

The average path length is still the highest for the nebo. Apparently the paths in these subtrees are, for the nebo and the nblo, relatively long because the average path length for each subtree is, with one small exception, i.e. *Road vehicles* for nebo, higher than the average path length for both the ontologies as shown in Table 7.1. The average path length for the pbo is higher than for each subtree.

The average number of subclasses for the pbo is higher than for the subtrees mentioned in Table 7.3. For the nebo the average number of subclasses is higher in the subtrees for *Road vehicles* and *Water vehicles* but smaller in *Buildings*. For the nblo only the subtree *Road vehicles* has a higher average number of subclasses than that number for the entire ontology.

With respect to entropy the subtrees of the nebo have the highest value compared with the subtrees of the pbo and nblo. The subtrees for *Buildings* of the nebo and the nblo show a remarkable higher entropy than for other subtrees. The correlation between entropy for the subtrees of all three ontologies and the number of concepts is strong: $r(7) = .97, p < .001$.

7.3 First Experiment

To find out which of the three ontologies is best suited for usage in SAQG we first conducted a pilot experiment. The method and results are presented in respectively Section 7.3.1 and Section 7.3.2. A discussion of the method and results is presented in Section 7.3.3.

7.3.1 Method

The experiment was done with 20 participants who tested the three ontologies. These ontologies were used to determine nine different situations shown in nine videos. The videos were downloaded from social media sites such as YouTube. The videos were prepared by cutting them all to a length of two minutes and disabling sound. The participants used the application, SAQG, on their own mobile device. All participants were students of the Amsterdam University of Applied Sciences and between 18 and 22 years of age.

Experimental procedure

Participants were gathered in one room and given an instruction on the experiment. During this instruction they were told what the sequence of events was going to be. First they downloaded SAQG and ontologies from a server to their mobile device. The participants were explicitly told that when an answer was chosen they were not allowed to return on an already given answer: they could not reconsider earlier choices and choose another path (in terms of Graphical User Interface: it was not allowed to navigate backwards). This instruction was given to make the simulation of a conversation by phone, which is done at emergency call centres, as close to reality as possible.

One of the first screens shown in SAQG was a screen where the participant had to choose an ontology. The first ontology to choose, was the pbo. Then a video was started about a building on fire. After a minute they were allowed to start SAQG and answer the questions. After two minutes the video stopped. When all participants had chosen a concept which according to them was the best representation of the object on fire, they chose in SAQG to finish the session and were directed to the start of a new session. Then they chose the nebo and the next video about a building on fire was started. The last ontology used in this series was the nblo. When the series about buildings on fire was finished, the series about water vehicles on fire was started with the same sequence of ontologies to use. Finally the three videos about a road vehicle on fire were shown. The sequence of the nine videos observed by the participants is shown in Table 7.4.

Presented video	**Ontology**
Building 1	pbo
Building 2	nebo
Building 3	nblo
Water vehicle 1	pbo
Water vehicle 2	nebo
Water vehicle 3	nblo
Road vehicle 1	pbo
Road vehicle 2	nebo
Road vehicle 3	nblo

Table 7.4: The sequence of observed videos and used ontologies.

Metrics

We logged each question presented, all the presented answers to choose from, each answer given and the time needed to give an answer. Because we were interested in how fast an object was determined we computed the average of all the times each individual participant using a particular ontology needed to determine an object. The correlation, i.e. the Pearson's product moment correlation coefficient, between the time needed to determine an object and the characteristics of each ontology, was also computed.

To answer the question whether and to what extent path length has influence on the time needed to determine an object on fire we divided the time needed to determine an object by the path length of the subtree used. For the number of subclasses we did the same. A correlation of the results of this metric with the time needed to determine an object was computed.

Of course we were also interested in what concepts were chosen by the participants as a representation of the objects on fire and how fast these concepts were chosen. Therefore we computed the variability, i.e. the number of different concepts chosen by the participants, and relative mode, i.e. the frequency of the most chosen concept relative to the total number of choices, for each presented situation and ontology. The ontology best suited for the task at hand has a small variability and a high relative mode.

7.3.2 Results and Analysis

The time needed to determine an object on fire is shown in Table 7.5. It can be seen that participants determine an object on fire in the fastest way when using the pbo. When using the nebo it takes them nearly two times as long. Also, it can be seen that the determination of a building on fire takes nearly the same time when using the nblo as when using the nebo.

When comparing Table 7.5 with the entropy as shown in Table 7.3 it can be seen that the correlation is strong and significant: $r(7) = .87, p = .0025$. The correlation with the number of concepts is a bit weaker: $r(7) = .83, p = .0059$.

To find out whether there is a relation between the path length and the time needed to determine an object we compared these values. It turns out that

Video	pbo	nebo	nblo
Buildings	23.52 (08.27)	49.07 (17.59)	43.45 (12.10)
Water vehicles	26.56 (09.18)	33.36 (11.39)	22.98 (05.39)
Road vehicles	17.98 (04.88)	35.40 (14.09)	20.31 (10.95)
Mean	22.68	39.61	28.91

Table 7.5: Mean time needed to determine an object on fire in experiment 1.

the correlation between the path length of each subtree and the time needed to determine an object is not significant: $r(7) = .59, n = 9, p = .098$. A correlation could also not be found when the time needed to determine an object was corrected with the path length. As can be seen in Table 7.6 when correcting the time needed to determine an object on fire for the average path length of the subtree, i.e. time divided by average path length of the subtree, the ontology which generates in the fastest way a determination of an object on fire is the nblo. A correlation between the time corrected by the path length and the time needed to determine an object on fire is not significant: $r = .47, n = 9, p = .2$.

Video	pbo	nebo	nblo
Buildings	13.75	13.33	15.97
Water vehicles	14.44	8.18	7.56
Road vehicles	10.52	9.86	7.05
Mean	12.90	10.46	10.20

Table 7.6: Mean time needed to determine an object on fire corrected by the average path length of the subtree in experiment 1.

When comparing the average time needed to determine the object on fire corrected by the path length in Table 7.6 with the average number of subclasses of the subtree in Table 7.3, it shows that the pbo has the highest average number of subclasses and the nblo the smallest (except for the subtree 'Buildings' but then in Table 7.6 the value for nblo is also higher than the value for nebo for the subtree 'Buildings'). But the difference in the average number of subclasses, nearly twice as much for the pbo as for the nblo, is not as strong as the difference in time needed as shown in Table 7.6. A correlation between these values is not significant: $r = .35, n = 9, p = .36$.

Results, when correcting the time needed to determine an object on fire with the average number of subclasses of the subtrees, are shown in Table 7.7. The correlation with the time needed to determine an object on fire is rather strong: $r = .90, n = 9, p = .001$. The correlation between the time needed to determine an object on fire and the number of subclasses of each subtree is not significant: $r = .37, n = 9, p = .33$.

The participants used various concepts to determine the object on fire as can be seen in Table 7.8. The number of different concepts chosen by the participants is shown (#) and the number of concepts most frequently chosen relative to the total number of concepts chosen, i.e. the mode relative to the total number of choices, is given (%). For example: In the video for *Building* for pbo two concepts were chosen ('Residence' and 'Residence Building'). The concept 'Residence' was chosen by 17 participants, which is 85% of the total number of concepts chosen.

Video	pbo	nebo	nblo
Buildings	4.41	15.05	12.21
Water vehicles	3.65	7.58	8.36
Road vehicles	3.37	7.71	5.30
Mean	3.81	10.12	8.62

Table 7.7: Average time needed to determine an object on fire corrected by the number of subclasses of the subtree in experiment 1.

Subject	pbo		nebo		nblo	
	#	%	#	%	#	%
Buildings	2	85	9	25	8	40
Water vehicles	5	50	9	30	3	80
Road vehicles	4	60	11	15	2	65
Mean	3.66	65	9.66	22.33	4.33	61.6

Table 7.8: Variability and relative mode in experiment 1.

The use of the nebo on average does generate more different answers from the participants than the use of the pbo and nblo, as can be seen in Table 7.8. Only when watching the video about a building on fire and using the nblo generates nearly the same amount of different answers. The specific video used in combination with the nblo shows a multi-story building which is an apartment building but could easily be mistaken for an (old) office building, department store (shops are on the first floor in buildings surrounding the building on fire) or a monumental building. Overall we can conclude that the greater number of concepts to choose from when using nebo, led to a less focussed choice of all the participants than when using the pbo and nblo.

7.3.3 Discussion and Conclusion

It is clear that an object on fire is determined in the fastest way when using the pbo and it takes the longest time when using the nebo. We can conclude that nebo has the highest variability and the lowest relative mode. The pbo has the lowest variability and the highest relative mode, although the difference with nblo is small. The correlation of the time needed to determine an object on fire and the entropy of the subtree of the ontology is rather strong. The nblo always has results in between the pbo and nebo except for the mean time to determine an object corrected by the average path length.

The influence of the path length and number of subclasses is not so manifest. The correlation of the time needed to determine an object corrected by the number of subclasses with the time needed to determine an object is strong. But the correlation of the time needed to determine an object with the path length is not significant.

Methodological considerations

This pilot showed that SAQG, the videos and other elements of the experiment were the right ingredients for doing research on this subject. Some improvements on the instruction and procedure would make the experiment better and the validation of the ontologies stronger.

The first thing to improve was to omit the instruction which allowed only to choose from what was presented as possible answers and not to return to the previous set of answers. This instruction led to some odd answers such as 'Urinal' when the video about a road vehicle on fire was shown with the nebo ontology. Reason for this specific inappropriate answer was the erroneous choice for 'Street object' after the first question and no backward choice was allowed.

The second improvement was found in the difficulty to compare the answers from the questions generated from the three ontologies. Each ontology was tested with a different video. All these videos showed different objects on fire and some of these objects on fire were rather difficult to determine. This experimental arrangement has lead to an undetermined bias.

The third improvement we decided to make, was the use of one type of mobile device because in this experiment participants used their own mobile device. This led to a variety of screen sizes. Most of the screens were more or less the same size but some participants used a mobile device which was considerable larger than average such as a tablet. The application showed these participants the same list with possible answers to a question as other participants but because their screens were larger they could more easily see a larger part of the list with possible answers. Because we did not take the specific size of the screen of the mobile device of the participant into account, these differences in screen size possibly could lead to differences in individual results.

The fourth improvement was to gather additional data on the time needed to determine an object. The number of subclasses did not affect the time needed to determine an object. When the time needed to determine an object was corrected by the number of subclasses a correlation of this corrected time with the time needed to determine an object was detected. But on the other side, the path length did not have a correlation with the time needed to determine an object.

We conclude that although a tendency in the data can be discerned, the design of the experiment created too much noise for clear results. Therefore, a new experiment was conducted.

7.4 Second Experiment

For the second experiment we implemented the improvements as discussed in Section 7.3.3. The method used in this experiment is presented in Section 7.4.1. The result and analysis are presented in Section 7.4.2.

7.4.1 Method

The experiment was done in eight sessions with a total of 110 participants. The smallest session was done with 5 participants, the largest with 23 participants. All participants were students of the Amsterdam University of Applied Sciences and between 18 and 22 years of age.

Experimental procedure

We used three videos instead of nine as in the previous experiment to reduce the noise which prevents a clear result (see Section 7.3.3). Each video was about an object on fire in the same domains as chosen in the previous experiment: *Buildings*, *Water vehicles* and *Road vehicles*. All the videos were cut to one minute and stripped of sound.

Each participant only used one ontology while observing the successively presented situations to prevent interference between the use of different ontologies. The videos were shown to all the participants while each of them was using one of the three ontologies. The participants did not choose the ontology to use, the application determined which ontology was used.

The first time a particular ontology is used a participant does need more time to determine the situation than the second and third time using the same ontology but determining different situations. To avoid such a training effect we started with a video_0, a video which showed a situation and was determined by the participants but was not taken into account for the experiment. This video showed a situation with an object on fire which easily could be determined as *Aircraft*.

For instruction we used a sheet which presented the participants the goal of the experiment, what an ontology is, and the sequence of steps to make the application work, e.g. making connection to the Internet. During the instruction particular attention was given to the possibility of backward navigation, i.e. return to a previous question and the possibility of choosing a superordinate concept when the subordinate concepts are not known or suitable.

All the participants were using the same mobile device: an IDEOS X5 with Android 2.2.1. Because the mobile device we used was the same for all participants and most probably different from their own, the participants were given a small amount of time to get used to the mobile device. After the participants entered their participation number an ontology was automatically chosen. The first question then was 'What is on fire' with the same multiple choices for all the ontologies (see Figure 3.4).

Metrics for the time needed to determine an object

Just as in the first experiment we logged the data to gather results and create statistics. We also compared the time needed to determine an object with the path length and number of subclasses and corrected the time needed to detect an object with these two characteristics.

To gather additional information we developed a questionnaire. After each video we let the participants answer some questions about the application. The questions were 'Do you miss a concept which describes the object on fire better?', 'What do you think of the sequence of questions?', 'Do you understand all the used concepts?' and 'What do you think of the Graphical User Interface?'. The first question could be answered by yes or no. The other questions could be answered by choosing from a scale of 1 to 5 a number which reflected their evaluation on this topic, where 1 was a negative evaluation and 5 a positive evaluation.

Metrics for the chosen concept

To determine whether a participant chose the right concept for the object on fire which was shown on the video, we developed three metrics.

	Preferred	**pbo**	**nebo**	**nblo**
Buildings	bungalow			
Preferred		woning	bungalow	bungalow
Alternatives		woongebouw	woonhuizen	woning
Water vehicles	cruiseschip			
Preferred		cruiseschip	passagierschepen	cruiseschip
Alternatives		passagierschip	zeeschepen	passagierschip
		ferry	veerboot	zeeschip
Road vehicles	schoolbus			
Preferred		bus	schoolbus	bus
Alternatives		-	Bussen_voertuigen	-

Table 7.9: Gold standard and alternatives. The preferred concepts as chosen by knowledge experts independent from the ontologies, concepts drawn from the ontologies close to the preferred concepts and alternative concepts drawn from the ontologies.

For the first metric we created a gold standard. Three knowledge experts observed, independent from each other, the presented situations and chose the concepts they thought best represented the object on fire shown on each video independent of any ontology. After a discussion there was agreement on the concepts which are shown in Table 7.9 in the column 'Preferred'.

For the second metric a concept from each ontology was drawn. This concept was similar to the preferred concept. The three knowledge experts also chose some concepts closely representing the object on fire, i.e. alternatives. In a discussion the perfect answer and the alternative concepts were determined. To compare the results on this metric per ontology each answer was given a value: the accurate answer was given one point and the alternatives half a point. These concepts are shown in Table 7.9 under the respective ontologies.

For the third metric it was determined how many different concepts were chosen by the participants, i.e. the variability of the chosen concepts, and how often the most chosen concept was chosen relative to all other choices, i.e. the relative frequency of the most often chosen concept. We expect consensus among the participants about the object on fire for each ontology used after watching the three videos. The level of agreement among the participants may also indicates the suitability of the ontology for this task.

7.4.2 Results and Analysis

How long it took for the participants to determine the object on fire is shown in Table 7.10. The participants which used the nebo on average needed much more time than the participants which used the pbo and the nblo. The difference between the time needed to determine the object on fire when using the pbo or the nblo was not significant (two-sample $t(210) = .60, p = .53$). The correlation between the time needed to determine the object on fire and the entropy of the subtree as shown in Table 7.3 is less than this value in experiment 1 but still significant: $r = .73, n = 9, p < .05$. The correlation with the number of concepts is even a bit stronger: $r = .77, n = 9, p < .05$

Table 7.11 shows how the ontologies relate to each other when the mean duration of determination for each video is corrected by the average path length of the subtree which was used. The differences between the ontologies are stronger than the differences between the ontologies in experiment 1 (see Table 7.6).

	pbo N=36	**nebo** N=35	**nblo** N=39
Video 1	29.54(20.97)	64.73(26.59)	40.49(16.72)
Video 2	43.55(34.75)	80.59(39.53)	37.97(31.65)
Video 3	25.97(16.35)	34.25(21.15)	26.77(10.48)
Mean	33.02(26.16)	59.86(35.52)	35.08(22.17)

Table 7.10: Mean duration of determination and standard deviation for each video and ontology in experiment 2.

The correlation of the time needed to determine an object corrected by the path length with the time needed to determine an object on fire is not significant. The correlation between the path length as shown in Table 7.3 and the time needed to determine an object is significant: $r = .68, n = 9, p < .05$. The high standard deviation is caused by outliers which sometimes occur.

	pbo	**nebo**	**nblo**
Video 1	17.28(21.05)	17.47(18.78)	14.89(15.62)
Video 2	24.61(30.72)	19.72(21.79)	12.49(16.12)
Video 3	15.19(17.43)	9.60(11.21)	9.30(9.88)
Mean	19.02(23.07)	15.59(17.26)	12.22(10.54)

Table 7.11: Mean duration of determination for each video and ontology corrected by the average path length of the subtree in experiment 2.

The time needed to determine the object on fire corrected by the number of subclasses is shown in Table 7.12. The correlation of the time needed to determine an object corrected by the number of subclasses with the time needed to determine an object on fire is rather high: $r = .86, n = 9, p < .05$. The correlation of the number of subclasses of a subtree and the time needed to determine an object using that subtree is not significant: $r = .17, n = 9, p = .66$.

	pbo	**nebo**	**nblo**
Video 1	5.54(6.75)	19.72(21.20)	11.37(11.94)
Video 2	6.22(7.77)	18.28(20.21)	13.81(17.82)
Video 3	4.87(5.59)	7.51(8.77)	6.99(7.43)
Mean	5.56(6.70)	15.17(16.73)	10.72(12.40)

Table 7.12: Mean duration of determination for each video and ontology corrected by the number of subclasses of the subtree in experiment 2.

The Tables 7.13, 7.14 and 7.15 show what choices were made by the participants. These tables show that participants chose many more different concepts to determine the object on fire when using the nebo than when using the pbo and nblo. When using the pbo the participants were nearly unanimous when determining a building on fire in video 1 and totally unanimous when determining a road vehicle on fire. When seeing a building and a road vehicle on fire when using the blo this also largely lead to agreement.

Table 7.16 shows the score for each ontology using the gold standard created independent of concepts which are part of the ontologies (see Section 7.4.1). For each time a participant chose the preferred concept (see Table 7.9) an ontology

pbo		**nebo**		**nblo**	
Woning	28	Woonhuizen	8	Woning	35
Woongebouw	8	Zomerhuisjes	7	Gebouw	1
		Bungalows	6	Vakantiehuis	1
		Stadshuizen	4	Bungalow	1
		Tuinhuizen	3	Pand	1
		Landhuizen	2		
		Krotten	1		
		Appartementen	1		
		Vakantieoorden	1		
		Seizoenwoningen	1		
		Woningen	1		

Table 7.13: Choices made by the participants using the pbo, nebo and nblo after seeing video 1.

pbo		**nebo**		**nblo**	
Cruiseschip	14	Zeeschepen	13	Cruiseschip	27
Passagiersschip	13	Schepen	9	Zeeschip	5
Ferry	4	Rivierschepen	2	Veerboot	4
Scheepvaartuig	3	Plezierjachten	2	Passagiersschip	2
Zeeschip	1	Boten	2	Passagiers of goederenschip	1
Rondvaartboot	1	Motorboten	1		
		Visserschepen	1		
		Kanaalschepen	1		
		Marineschepen	1		
		Vaartuigen	1		
		Stoomschepen	1		
		Pleziervaartuigen	1		

Table 7.14: Choices made by the participants using the pbo, nebo and nblo after seeing video 2.

pbo		**nebo**		**nblo**	
Bus	36	Schoolbussen	28	Bus	37
		Andere voertuigen	4	Hulpdienst	2
		Recreatievoertuigen	2		
		Wagentje	1		

Table 7.15: Choices made by the participants using the pbo, nebo and nblo after seeing video 3.

	pbo	**nebo**	**nblo**
Video 1	0	6	0.90
Video 2	13.61	0	24.23
Video 3	0	29	0
Total	13.61	35	25.13

Table 7.16: Validation of the ontologies using the gold standard independent of ontologies (normalized).

scores a point and when an alternative was chosen half a point was given. To compare the ontologies a normalization has been applied because the number of participants using the pbo was 36, the nebo 35 and the nblo 39, the result was corrected by respectively $\frac{35}{36}$, 1 and $\frac{35}{39}$. It is clear that nebo scores best and pbo worst using this metric. For participants that used the pbo it was not possible to choose the right concept when seeing video 1 and 3. The same holds for the nebo with video 2 and the nblo with video 3. The concept 'Cruiseschip' was missed by 11 participants (see Table 7.20). This concept was element of the original ebo and erroneously omitted in the conversion to nebo. When this is taken into account the score for nebo even becomes better. The participants which used the nblo with video 1 who wanted to choose 'Bungalow', had to choose 'Holiday accommodation' ('vakantieverblijf') first instead of 'Residence' ('woning'). Figure 7.2 shows the subtree for 'Buildings' ('gebouwen'). As can be seen in Table 7.13 most participants chose 'Residence' ('woning').

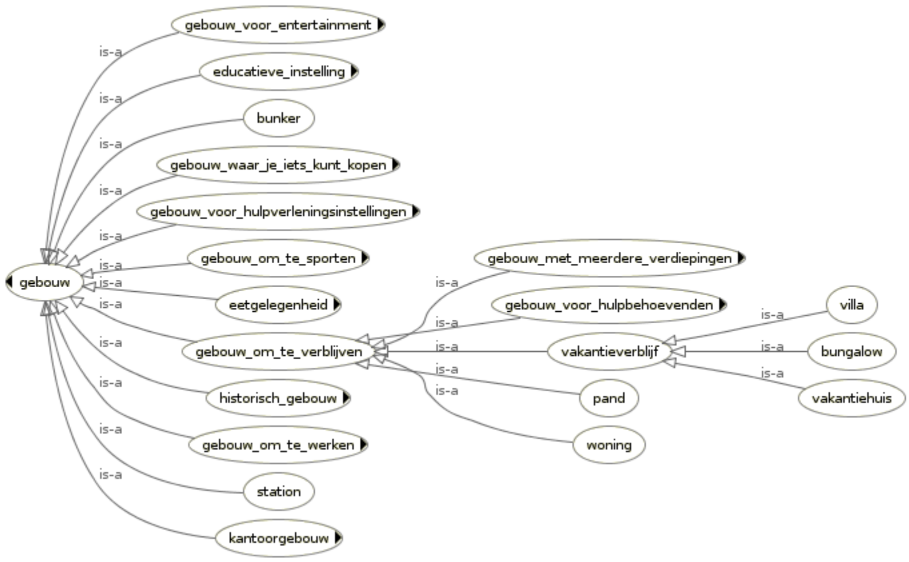


Figure 7.2: Subtree Buildings of the new basic-level ontology

	pbo	nebo	nblo
Video 1	31.11	10.50	16.71
Video 2	21.88	6.50	27.37
Video 3	35.00	29.00	33.21
Total	87.99	46.0	77.29

Table 7.17: Validation of the ontologies using the gold standard and alternative concepts retrieved from ontologies (normalized).

Table 7.17 shows the score for each ontology using the gold standard and alternatives created with use of concepts that are part of the ontologies. The lowest score is for nebo. The best score is for pbo. Due to the strong variability of choices made by participants using the nebo, this ontology does not score high on video 1 and 2. Only with video 3 the nebo scores nearly as high as pbo and nblo. The variability of choices for all ontologies after seeing video 3 is much smaller as can be seen in Table 7.15.

Table 7.18 shows how many different choices were made by the participants

Subject	pbo		nebo		nblo	
	#	%	#	%	#	%
Buildings	2	77.78	11	22.86	5	89.74
Water vehicles	6	38.89	12	37.14	5	69.23
Road vehicles	1	100.00	4	80.00	2	94.87
Mean	3.00	72.22	9.00	46.67	4.00	84.61

Table 7.18: Variability (#) and relative frequency of the most often chosen concept (%) in experiment 2.

and what percentage of the total the most frequently chosen concept is, i.e. the mode relative to the total number of choices. For this metric the best score is for nblo and the worst score for nebo. For the video about a road vehicle the users of the pbo unanimously chose for 'Bus'. The route taken for participants using the pbo is shown in Figure 7.3.

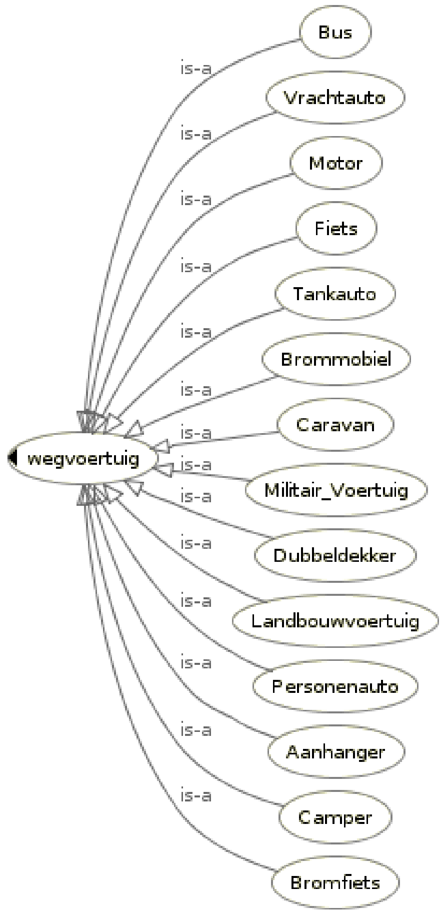


Figure 7.3: Subtree Road vehicles for the pragmatic-based ontology.

In Table 7.19 the results from the questionnaire are shown. The participants did not miss concepts in the determination process in one ontology more than in another. Taken all ontologies together nearly 63% of the participants did not miss a concept to describe the object on fire. The pbo, which is the smallest ontology, scores worst on this question.

		1	2	3	4
pbo	mean	0.59	4.22	4.15	2.83
n = 36	video 1	0.611	4.36	4.36	2.83
	video 2	0.81	3.94	3.44	2.83
	video 3	0.36	4.36	4.64	2.83
nebo	mean	0.67	4.03	3.51	3.26
n = 35	video 1	0.69	4.20	3.40	3.29
	video 2	0.49	3.57	3.03	3.17
	video 3	0.83	4.31	4.11	3.31
nblo	mean	0.68	4.30	4.44	3.29
n = 39	video 1	0.72	4.23	4.36	3.28
	video 2	0.90	4.36	4.26	3.28
	video 3	0.41	4.31	4.69	3.31

Table 7.19: Results from questionnaire. 1: 'Do you miss a concept which describes the object on fire better?' (no(1)/yes(0)), 2: 'What do you think of the sequence of questions?' (1-5), 3: 'Do you understand all the used concepts?' (1-5), 4: 'What do you think of the Graphical User Interface?' (1-5).

Comparing the results of the questionnaire for each ontology we see a difference in mean for the comprehension of concepts between the nebo (3.51) and nblo (4.44) which is significant ($F = 5.77, p < .05$). The difference in mean for the Graphical User Interface between the pbo (2.83) and nebo (3.26) is significant ($F = 4.71, p < .05$).

When we compare the results for the time needed to determine an object (see in Table 7.10) with the results on the question whether the participant understands the concepts a significant correlation occurs: $r = .91, n = 9, p < .001$. When comparing the time needed to determine an object with the results on the evaluation of the sequence of questions also a significant correlation can be found: $r = 0.83, n = 9, p =< .05$. The correlation between the two questions just mentioned is also significant: $r = .84, n = 9, p < .05$. A comparison between the values for the question after the Graphical User Interface and the average number of subclasses for each subtree shows a significant correlation: $r = .84, n = 9, p < .05$.

	pbo		nebo		nblo	
Video 1	Alleenstaand	6	Alleenstaand	4	Alleenstaand	3
	Bungalow	2	Eengezinswoning	3	Bungalow*	3
	Vakantiewoning	2	Off task	4	Huis	2
	Woonhuis	1			Off task	3
Video 2	Cruiseschip*	1	Cruiseschip*	11	Zeeschip*	1
	Passagiersschip	1	Ferry*	3	Veerpont	1
	Off task	4	Off task	4	Off task	2
Video 3	Schoolbus	19	Schoolbus*	3	Schoolbus	20
	Personenvoertuig	1	Bus	1	Bus*	1
	Off task	3	Off task	2	Off task	2
Total		40		24 (+11)		38

Table 7.20: Concepts missed by the participants per ontology and per video ('*' indicates it is a concept that participants failed to notice in the ontology).

In Table 7.20 the concepts missed by the participants are shown. Each video and ontology has a category ‘Off task’, containing answers such as ‘type of material’ or ‘show some more of the surroundings’ which are not relevant for this research. On video 3 there is a remarkable high score for pbo and nblo on the concept ‘Schoolbus’. The low value for this video when using pbo or nblo, which can be seen in Table 7.19, is accompanied by the spontaneous mentioning of the concept that was indeed the preferred concept as given by the knowledge experts (see Table 7.9). The same can be said about the missing concept in nebo after seeing video 2, i.e. ‘Cruise ship’.

7.5 Discussion

The methodological adjustments to the method of the experiment have led to a more clear-cut result. Odd choices were not made and all the participants were able to use backward navigation. All three ontologies were evaluated with the same situation which prevented a bias. The mean time to determine an object on fire shows the same relative order in both experiments (comparing Table 7.5 and 7.10). Whether the path length or the number of subclasses influences the time to determine an object was not decided in the first experiment but is clarified in the second experiment. Also, by using a questionnaire we found other aspects of the ontologies that influence the time to determine an object. The variability and relative mode show nearly a similar result: only the relative mode for ‘Building’ when using the nblo is different in the experiments.

We found five characteristics of the ontologies that have an influence on the time needed to determine an object: path length, comprehension of concepts, sequence of questions, number of concepts and entropy. The path length of an ontology has a direct influence on the time needed to determine an object because it is a metric reflecting the number of questions asked. Whether participants comprehend concepts has a less direct influence and is probably dependent on the knowledge of the participant which can vary. The reason why the sequence of questions has influence is less clear. It might be a derivative of the number of questions and thus the path length. The number of concepts and entropy have a strong correlation although the entropy metric also incorporates path length and number of subclasses. Measuring entropy of an ontology has not been exhaustively researched until now, as has been discussed in Section 6.6.

The analysis of the choices made by the participants which was done using a gold standard shows rather confusing results. While using a gold standard set up by knowledge experts independent of the ontologies shows a best score for nebo and using a gold standard of concepts drawn from the ontologies shows a best score for pbo. To make matters even more confusing, an analysis of the choices made by participants based on the number of participants choosing the most popular concept relative to the total number of participants shows the best score for nblo.

The gold standard independent of the ontologies is, to our opinion, a metric for evaluating the completeness of an ontology, i.e. the number of concepts from an ontology also found in a corpus relative to the total number of concepts in that corpus, i.c. natural language. This experiment makes clear the difference of determination of an object on fire by knowledge engineers on the one side and ordinary people on the other. The gold standard drawn from the ontologies does reflect the suitability of the ontologies for use by ordinary people much better. The outcome would even be similar to the outcome of Table 7.18, which shows the consensus among the participants, when for video 1 (Building) the gold standard would have been 'Residence' (Woning) instead of 'Bungalow'. The choice for 'Bungalow' was again due to our own knowledge engineering perspective. That this expert view does comply to a certain degree with natural use of language is also clear from the choice of preferred and alternative concepts. When the distinction between preferred and alternative concepts was not as strict as in our metric the nebo would have scored better. Erroneously omitting the concept 'Cruiseship' from nebo did not have effect on the overall results. The used metrics, i.e. the number of different choices (variability) and the relative mode, as is shown in Table 7.18 is, in our opinion, most informative about the suitability of the ontologies for the task we envisage.

An indication of completeness of the ontologies was given by the answers to the question for the missed concepts we stated in the questionnaire. The participants missed the most concepts in the pbo and only slightly less in the nblo. In the nebo they missed considerably less concepts when the erroneously omitted concept 'Cruiseship' is not taken into account. The conclusion that nebo is more complete than the pbo and nblo can be explained by the larger number of concepts of the nebo (see Tables 7.1 and 7.3). When the number of concepts is taken into account one should conclude that the specificity of the nebo is low. This could also explain the high variability in the answers for the nebo: when participants have more choice they give more different answers.

7.6 Conclusion

In this chapter two experiments are presented that were conducted to determine whether one of the three ontologies is most suitable for the task of answering by ordinary people of automatically generating questions for determination of a situation. The three ontologies were each developed in a different way as described in Chapter 6. In Chapter 6 we measured in several ways the suitability of the ontologies. The experiments presented in this chapter were also used to validate the measurement done in Chapter 6. Before we could use the ontologies, two of the ontologies had to be re-engineered. Then we did a pilot experiment. A second experiment was done after some methodological adjustments.

The participants do understand the concepts used in the pragmatic-based ontology but evaluate the Graphical User Interface, relative to the nebo, negative. This is probably due to the large number of subclasses of the pbo. The pbo scores best on the time needed to determine an object on fire (but the difference with the time needed when using the nblo is not significant). Using the gold standard independent of the ontologies the pbo scores worst, but when using the gold standard with concepts drawn from the ontologies the pbo scores best. The pbo shows the smallest variability of choices when determining an object on fire. The average relative mode of the pbo is rather high but not the highest.

The concepts used in the new expert-based ontology are the least understood relative to the nblo. Using the nebo it took the participants the most time to determine an object on fire. Using the gold standard independent of the ontologies the nebo scores best, but when the concepts are drawn from the ontologies pbo and nblo score much better and nebo scores worst. The nebo shows the largest variability and the lowest relative mode. The greater number of concepts for the nebo seems to be the cause of this result.

The participants understand the concepts used in the new basic-level ontology rather well. When comparing the time needed to determine an object on fire the nblo scores nearly as good as the highest scoring ontology. The gold standard metric for nblo shows a mediocre result. The variability is rather low but not the lowest. The average relative mode is the highest of the three ontologies.

When predicting the time needed to determine an object on fire the average path length and the number of concepts used in an ontology are good indicators. Path length has a rather strong correlation with the time needed to determine an object. The number of subclasses does not have a correlation with the time needed to determine an object. Comprehension of the concepts also has a strong influence on the time needed to determine an object but is dependent on the knowledge of users.

The measurements we did in Chapter 6 are validated by the results from this experiment. The structure and cognitive semantics of the basic-level ontology were evaluated as being good or even very good (see Table 6.8) are in agreement with the results of this experiment. For the expert-based ontology the results of these measurements were also confirmed.

We conclude that each of the ontologies has its own merits and idiosyncrasies. When an accurate determination of an object is needed and users are knowledgeable, one should use the nebo. A longer time to determine such an object has to be accepted. The nblo scores nearly as good or better than the pbo in most respects, such as time needed to determine an object, variability and consensus about the concept that denotes the object on fire best. The nblo is evaluated more positive than other ontologies with respect to the user interface. A great disadvantage is the high costs to develop a nblo (see Chapter 6). When an application such as SAQG is deployed on a large scale for many people such an effort could be worthwhile. The concepts and their categorization represented in the ontology used by SAQG to generate questions and possible answers should comply maximally with the language used by ordinary people. The new basic-level ontology has the best results in this respect.

Chapter 8

Conclusion and Discussion

8.1 Conclusions

This thesis presents research on the possibility of gathering information from a large group of ordinary people about a crisis situation. In a crisis situation it is paramount to determine the situation as fast as possible. The research question we answered is: 'Is it possible to automatically generate questions which can be answered in such a way by ordinary people that the nature of a crisis situation can be determined in a reliable way?'. This question is decomposed into several research questions (see Chapter 1) that are answered here. We conclude that the gathering of information fast and from a lot of people can be done within the framework we developed and with the help of an application as presented before. During this research several observations gave rise to discussions (Section 8.2) which could lead to further research (Section 8.3). In Section 8.4 some concluding remarks are presented.

RQ I. What is an appropriate framework for the formalization of crisis situations?

Situation Theory is used to create a framework which formalizes information and situations. With Situation Theory a situation is described as a construction of smaller informational entities, i.e. infons. Parameters and types make it possible to represent specific sets of situations and other informational entities and to create abstractions of real world situations and objects. This abstraction is a fundamental method to generate questions. Furthermore, constraints on situations and infons represent rules such as necessity, causality and social obligations. To represent this formalization in a computer-readable format the web ontology language OWL was chosen. To map Situation Theory on an ontology represented in OWL, we revised an existing ontology according to principles of knowledge engineering. With Situation Theory Ontology Revised (STOR) specific situations can be defined and questions automatically generated.

RQ II. Given a model to formalize situations, how can we use this model to generate relevant questions?

The Theory of Strongly Semantic Information is used to make a choice from all possible questions and pose the most informative one. Five strategies are defined to automatically generate the most informative questions to determine a situation. The first strategy is applied when no assumptions about the possible situations can be made and consists of subsequently asking questions about each possible situation. The second strategy is applied when all the possible situations are instances of the same set of parametrized infons and the abstraction is over a set of multiple referents. The choice of a particular referent leads to an additional specification of the situation. The third strategy is applied when the parametrized infons only differ from each other by their polarity. The fourth strategy is applied when dependencies within the domain under question can be determined. Because these dependencies exclude specific situations it becomes possible to preclude impossible situations as a result of the determination of a situation. The fifth strategy is applied when the referents of an abstraction have a subsumption relation with each other.

An application, Situation Awareness Question Generator, was developed to present the questions to the user on a mobile device. SAQG is a client-server application that sends a domain ontology from a server to the client. This domain ontology is created such that it maps to a representation and generic ontology already installed on the mobile device. Questions are retrieved from the ontology and answers are added that are, when a situation is determined, sent back to a server.

RQ III. What methods are suitable to create a domain specific ontology for crisis situations?

With an experiment it was shown that an ontology describing dispositions of people created by knowledge engineers was not suitable for retrieving trustworthy answers. Participants in this experiment did not give reliable answers because they did not fully adhere to the concepts defined by these experts. A subsequent experiment showed that an ontology created from concepts used by professionals in the domain of crisis management did generate trustworthy answers. To find the characteristics of an ontology that generates trustworthy answers we created ontologies with different distinguishing characteristics.

We created domain ontologies by applying three different methods. The first ontology we created was a pragmatic-based ontology by re-engineering a categorization in use at an emergency call center to determine a specific situation. The second ontology was an expert-based ontology by merging two existing ontologies that were created by knowledge engineers. The third ontology was a basic-level ontology by eliciting concepts and their attributes from ordinary people and use an algorithm to create an hierarchy of concepts. To formally evaluate these ontologies four criteria were used: *a*) the ontology must have a structure suitable for the task, *b*) the construction of the ontology must be efficient, *c*) the ontology must be complete and *d*) the ontology must have a 'fit' with human categorization. The structure of the ontologies was evaluated by measuring the number of concepts, maximum and average path length and number of subclasses and coherence. The efficiency of constructing the ontologies was measured by taking the time needed to create the ontologies. The completeness was evaluated by comparison of the ontologies with a corpus elicited from ordinary people and measuring coverage and precision. The 'fit' of the ontologies with human categorization is discussed in the next section.

The validation of these metrics, i.e. whether these measurements indicate the suitability for the task of automatically generating questions and answers, was done by an experiment with SAQG and the three ontologies. The ontologies were slightly adjusted to be suitable for SAQG. The results of this experiment confirmed the conclusion that was drawn from the formal evaluation. The expert-based ontology is the most efficient to construct but lacks compliance with human categorization. Because of its large average path length users need more time to determine a situation than when using another ontology. The pragmatic-based ontology complies better to human categorization but misses a lot of concepts in use by ordinary people. The basic-level ontology is the least efficient to construct but complies best with human categorization.

RQ IV. What methods can be used to evaluate the 'fit' of domain specific ontologies with human categorization?

Several metrics to evaluate the theoretical 'fit' of domain specific ontologies with human categorization were used: entropy, Ingve-Miller number and semantic distance validation. Entropy was used to measure the amount of information contained in an ontology. The Ingve-Miller number was used to tell whether the number of presented possible answers was too large to comprehend in a fast way. An existing method used in psychological research was applied to measure the compliance between the hierarchy of the ontology and the natural categorization humans use. This method we called 'semantic distance validation' and was used to evaluate the 'fit' of the ontologies with human categorization. The result of the measurement of entropy showed that the pragmatic-based ontology had the smallest entropy and the expert-based ontology scored highest. The basic-level ontology scored between the other ontologies. The measurement of the Ingve-Miller number showed that the pragmatic-based ontology scored most unfavourable while the other ontologies scored better. For the validation of the semantic distance the basic-level ontology scored best and the expert-based ontology worst.

The validation of the ontologies in an experiment using SAQG showed that the basic-level ontology scored best on accuracy and time needed for determination of a situation. This result complies with the score of the basic-level ontology on the Ingve-Miller number and validation of semantic distance. The expert-based ontology scored better than the pragmatic-based ontology on the Ingve-Miller number but worse than the pragmatic-based ontology on the validation of semantic distance. In the experiment to validate the semantic distance the expert-based ontology scored worst.

RQ V. What is the reliability of answers from ordinary people to automatically generated questions about situations?

The reliability of answers from ordinary people to questions about a crisis situation depends on the ontology that is used to automatically generate questions. Questions retrieved from the pragmatic-based ontology result in answers that are reliable but do not give as much information as questions retrieved from a basic-level ontology. An expert-based ontology results in the least reliable answers when compared with pragmatic-based and basic-level ontologies. The basic-level ontology results in the most reliable answers from ordinary people.

The reliability of answers is related to the familiarity people have with the concepts used in a question. The more familiar people are with concepts the more reliable the answers will be. An abundance of possible answers as in the expert-based ontology has a negative effect on the trustworthiness of the answers, i.e. people choose a lot of different answers and have difficulty to understand the concepts used. The basic-level ontology, which is developed using concepts retrieved from ordinary people and an algorithm to detect the basic-levels, generates questions that are much better understood and answers that are more reliable. This is a confirmation of the theory of Rosch (Rosch et al., 1976a) that states that some concepts are much more used by people than others.

8.2 Discussion

8.2.1 Situation Theory as a Formal Framework

We used Situation Theory as the formal framework to represent situations and generate questions. Situation Theory offers types that can be restricted to add concepts for objects. With an abstraction mechanism and using parameters we generate questions asking for the actual replacement of the parameters. Relations are defined to capture one or more abstract constituents of a statement creating a flexibility and structure needed to represent the real world. These characteristics make Situation Theory much more fine-grained to describe real-world objects than other formalisms. Situation Theory is much richer than shown by the use of it in this thesis. We used concepts and the reasoning-mechanism available in Situation Theory tailored to the task at hand. An extension of the use of Situation Theory and its representation in OWL is realisable and desirable when gathering more information about a crisis situation. Furthermore, Situation Theory is based on the 'open-world assumption', just as OWL. Other formal systems such as 'situation calculus', 'simple event model' or a relational database are not appropriate for several reasons.

Situation Calculus (McCarthy, 1963; McCarthy and Hayes, 1968; Levesque et al., 1998) is created to reason about actions, which is reflected in its concepts and reasoning. Based on specific information possible actions are computed and an action is initiated. Situation Calculus is developed to compute the best strategy and achieve a goal by a formal system. We are after descriptions of the world and generate questions to retrieve the best description. Furthermore, Situation Calculus is not based on the 'open world assumption'.

The Simple Event Model is a graph model for events and related concepts such as actors, places and time (Van Hage et al., 2011; van Hage and Ceolin, 2013). It is very useful as an overall ontology for distinct information sources with different types because its high level of abstraction and it is simple with a small number of concepts. Also, it focusses on dynamic events while we need representations of much more static situations. The Simple Event Model is represented in OWL which is based on the 'open world assumption'. The

Simple Event Model does not enforce any format of information which makes it useful for many different ontologies. For example, the format of infons and situations are defined in Situation Theory and not in the Simple Event Model. To generate questions we need a more uniform format of information and much more detailed descriptions than the Simple Event Model can offer.

Using a relational database for storing concepts and the SQL query language for retrieving information would deprive us of the inherent relations an ontology has. We use these relations for reasoning to generate questions. Also, a relational database is not flexible because it restricts us to the defined format of the tables in a database. Furthermore, a relational database is not based on the 'open world assumption'.

8.2.2 Criteria for Alignment with Human Categorization

We used three methods to validate alignment of ontologies with human categorization: entropy, Ingve-Miller number and semantic distance validation (see Chapter 6). An experiment was conducted to determine a situation with our framework using three ontologies (see Chapter 7). This experiment showed that the ontology that generated questions and answers for ordinary people in the most efficient way also did best on the measurements we have chosen.

With entropy we measured the amount of information of an ontology that was used to automatically generate questions to determine a situation. Entropy measurement is applied to the entire ontology that is used to generate questions. We suggested that it does not make sense offering users very large ontologies because it slows down the determination process. The downside of such a choice is less completeness of small ontologies as is shown in Chapter 6 and 7. It remains the question what the amount of information is, ordinary people can handle in an efficient way.

The measurement of the Ingve-Miller number for the subclasses which are shown as possible answers to questions does not correlate with the efficiency of answering the questions. While the time needed to determine a situation differs between the ontologies, the Ingve-Miller number did vary only to a small extent. Probably because this measurement uses the number of subclasses which also did not correlate with the efficiency of answering the questions. The use of this measurement did not add to characterizing the ontologies.

The validation of semantic distance is a new method to validate the alignment of an ontology with human categorization. We chose three concepts in a random way and when the semantic distance, i.e. the path length between concepts, was larger between one concept and the two other concepts than the distance between these two concepts, this triad was presented to participants of the experiment. We did not do research on the most optimal measurement. Still, the results of the validation of semantic distance for the different ontologies corresponded to results for the ontologies in the experiment using SAQG.

8.2.3 Entropy or Information Value

The value of information is difficult to determine. In this thesis two different conceptions of the value of information are deployed: information value and entropy. These two conceptions come from two programs in philosophy: the transcendental program, which is rooted in Kant's philosophy and the empiricist, program which has its roots in Hume, according to Adriaans (Adriaans, 2010). The first conception is defined by Floridi as the semantic distance of the description of the situation from the actual situation. An inaccurate description of the situation also has a value. We used this conception of value of information to define the strategies for generating the most informative question. Entropy is described as the probability of occurrence, which we used to measure the amount of information contained by an ontology.

To define a strategy which excludes impossible situations we could not use entropy because then questions for impossible situations have the maximum amount of information and are always asked first. A question for an impossible situation within the transcendental program does not have informational value and is automatically not asked. Furthermore, strategies only using questions on vacuous situations gradually come to the right description of the situation.

We measured entropy of an ontology by taking the ratio of the number of subclasses of a concept and the total number of concepts because it was one of the few measurements of entropy of ontologies available at this moment. Use of this metric was not satisfactory if only because it has a strong correlation with the number of concepts (see Chapter 7). The metric for measurement of entropy uses the hierarchical structure based on *is-a* relations in an ontology. An alternative measurement uses the amount of definitions in an ontology, i.e. language-related and domain-related relations between the concepts (Doran et al., 2009). Which of these metrics measures entropy of an ontology best, is still to decide. All these measurements do not evaluate the value of an ontology for humans but give an indication of the diversity of information represented by the ontology. When measuring the suitability of ontologies for use by humans one needs to relate the ontology to how humans think. Entropy is only related to the amount of information humans can handle.

To validate the information value of an ontology one has to relate the structure of the ontology with the structure of human categorization. Such a validation could be done by measuring the frequency of use of the terms in natural language. Another measurement is presented in this thesis as the validation of semantic distance.

For our purposes measuring the information value of an ontology is much more suitable than the measurement of entropy because we use the ontologies to generate questions that must be comprehensible by ordinary people.

8.2.4 Generating Questions from a Formalization

Five strategies to determine the most informative question were developed. These strategies differ from each other by the restrictions we detected in the ontologies. The restrictions are a result of how knowledge of the domain at hand is structured. We have created ontologies with subsumption, i.e. an *is-a* relation, and properties, i.e. a *has-a* relation. These strategies always start with the question to which the answer is vacuous but nevertheless generates information. We start, so to speak, always at the top of the ontology. The strategies are developed with the goal to get the most specific determination of the situation as possible, as fast as possible.

To decrease the time to determine a situation, an alternative strategy is to ask people to describe the situation and focus on keywords. Then search for that concept in the ontology and start questioning from there in the ontology. Such a 'jump-start' possibly decreases the time needed to determine the situation.

The reliability of the answers from the users of our system is not perfect. Although the goal of 100% reliability probably never will be met, an improvement of the reliability is possible. An improvement can be realized by fine-tuning the ontology, e.g. creating an ontology which even better complies with human categorization. Furthermore, the ongoing research for mapping techniques and the availability of ontologies, e.g. Linked Data (Bizer et al., 2009), will make the expert-based ontologies more suitable for use in SAQG.

A probably even greater improvement can be made when the answers of users are suggested to other users. For example, a choice between ones own answer and the answers of others can be presented. Our application gathers information from several users but does not generate any conclusion about the information gathered from all the users together. All the information gathered from users can be aggregated into a bayesian system as is done in Distributed Perception Networks (Maris and Pavlin, 2006).

The application we developed (SAQG) can be extended in several ways. For example, an auditive system can be added so users can hear the questions and possible answers and respond by talking (Gatt and Reiter, 2009). Whether such a technique is faster in determination of a situation is unclear. On the one hand giving an answer is probably faster because the user does not have to use his/her fingers to touch the screen. On the other hand the reading of answers is probably faster than listening to a list of subjects to choose from. Such a technique does not generate extra information.

Other questions, such as about the number of people involved in a crisis and about the crisis itself, are also of interest for emergency services. With small adjustments most of this information can easily be formalized and questions can automatically be generated.

8.2.5 What about Twitter?

From the experiment to gather concepts for the gold standard in Chapter 6 we learnt that when people are asked to describe the situation from scratch, this is mostly done in general terms. With the advent of social networks such as Twitter and Facebook a lot of research is done to retrieve information from people by scanning descriptions of situations. We expect that this way to gather information does not generate as much information as our system. This is a consequence of the lack of incentives to be more precise and thus give more information to such systems. When people are asked to describe what they see, most of the time, they use very general terms such as 'there is a building on fire' and do not add further information. Our system asks 'what kind of building?'. Such questions generate more information. Using the first comment on a crisis situation by people can be used in an extension of SAQG when these general remarks are scanned on keywords that can be used to generate questions for a more detailed description.

8.3 Future Work

While doing research for this thesis we derived a gold standard to evaluate the ontologies which were build (see Chapter 6). The standard was created using terms participants of the experiment used in observation statements describing videos that show objects on fire. This standard was not suitable for evaluation because only a very few terms used were found in the ontologies. Other terms were referring to other objects than the object on fire. This suggests that when people are asked to describe a crisis situation in their own words, less information is given than when people are asked questions as is done with SAQG. To prove this hypothesis a new experiment has to be conducted. In this experiment SAQG is competing with social media applications such as Twitter to get as much information from ordinary people as possible.

This thesis only investigates one structure of human knowledge, i.e. categorization. I investigated a small part of categorization: only the subsumption and property relation. The part-of is also much used and of interest for determination of a situation. Furthermore, categorization is a rich and important aspect of human thought but not the only one. Other structures such as causality and association are also interesting.

Categorization is language, culture and time dependent (Rosch et al., 1976a). Research after the specific categorization used by the group of users is a very practical and satisfying because application of the knowledge from that research is immediately useful.

8.4 Final Remarks

In this thesis we have developed various methods for ontology-based question generation to elicit crisis information from the general public. Evaluation experiments have shown a complex trade off between development efficiency and completeness of ontologies on the one hand and compliance with human information processing capabilities on the other hand. So, optimal method selection depends on the goals of the information gathering process.

The information gathered with the application developed in this thesis can be of great use for emergency services. Because the application can be distributed among a large group of people, the information gathering can be done simultaneously. Such a distributed gathering of information presents the emergency services with a much greater amount of information than can be gathered by an emergency call center when a crisis situation involving a large number of people occurs.

Summary

To mitigate the effects of a crisis it is important to gather information as fast as possible. Ordinary people involved in a crisis often have been neglected by emergency services as a source of information. Nowadays they are more and more regarded as the true 'first responders'. In this thesis a framework is proposed, to gather trustworthy information about a crisis situation from ordinary people. This framework has been implemented as an application: the Situation Awareness Question Generator (SAQG). SAQG automatically generates questions from an ontology and uses the answers to determine a situation. Because in this framework number, form and content of questions is determined by the specific ontology in use, several experiments have been conducted to characterize the ontology which is most suitable for this task.

The problem statement in this thesis is as follows: *Is it possible to automatically generate questions which can be answered in such a way by ordinary people that the nature of a crisis situation can be determined in a reliable way?* This general question is decomposed in several specific research questions:

1. What is an appropriate framework for the formalization of crisis situations?
2. Given a model to formalize situations, how can we use this model to generate relevant questions?
3. What methods are suitable for creating a domain specific ontology for crisis situations?
4. What methods can be used to evaluate the 'fit' of domain specific ontologies with human categorization?
5. What is the reliability of answers from ordinary people to automatically generated questions about situations?

To formalize situations we used Situation Theory which has a richer structure than related formalisms and thus is much more fine-grained. In Situation Theory situations are represented by (compound) infons. By using parameters, infons can abstract over several possible situations in the real world. We mapped Situation Theory to OWL and created an ontology. For this we used Situation Theory Ontology and revised it, to make it more compliant with knowledge engineering principles.

We developed several strategies to automatically generate questions. Characteristics of an ontology determine which strategy to use. When the infons that define a situation are not related a rather laborious strategy is to be used.

When a situation is defined by parametrized infons it depends on the type of parameter which strategy is to be used. When rules are important another strategy is to be used. An ontology consisting of subsumption relations is being handled by yet another strategy. These strategies are implemented in SAQG which uses an algorithm to choose a strategy.

For an experiment to determine the disposition of people involved in a crisis situation we developed an ontology which determines different roles (Victim, Not-Active, Observer and Helper). Several scenarios, describing a crisis situation, were presented to participants. Having read these scenarios the participants answered questions. The answers were used to determine the disposition of the participants. The experiment showed that our ontology did not generate the questions which lead to the determination of dispositions as we envisaged. We concluded that our ontology did not comply with how people categorize possible dispositions during a crisis situation.

To create a different type of ontology, this time of car accidents, we used some concepts provided by experts mentioned in literature. We also retrieved concepts from two database in use by professionals in the field of car accidents. An experiment using SAQG showed that this ontology did generate questions which participants answered trustworthy, i.e. their answers corresponded to the actual situation.

To uncover the characteristics of an ontology suitable for automatic question generation we constructed three different ontologies. A pragmatic-based ontology was constructed from a vocabulary which the Amsterdam fire department uses to categorize calls for 112. An expert-based ontology was constructed from two ontologies developed by knowledge-engineers and experts in the field of art. A basic-level ontology was constructed from concepts and their attributes retrieved from ordinary people. Using the attributes we created an ontology with an algorithm based on the basic-level theory of Rosch.

We set up a framework to measure the suitability of these three ontologies which evaluated the ontologies on four aspects: *a*) the ontology must have a structure which is useful for automatic question generation, *b*) the construction of the ontology is efficient, *c*) the ontology must be complete and *d*) the ontology should be compliant with human categorization. For this framework we used several existing metrics such as maximum path length, number of concepts, entropy and the Ingve-Miller number. Also, we developed a new metric called 'semantic distance validation'. This metric compared the distance of concepts in terms of path length with how participants in an experiment evaluated this distance.

From these metrics we learnt that the expert-based ontology had a structure least suitable for automatic question generation because it has a large number of concepts, the longest path length, high maximum number of subclasses and a high entropy. The least efficient to construct was the basic-level ontology because of the laborious retrieval of concepts and attributes. The re-engineering of the vocabulary to create the pragmatic-based ontology did cost more time than applying an algorithm as was done to construct the expert-based ontology. The expert-based ontology is the most complete ontology and the pragmatic-based ontology is the least complete ontology. With respect to the compliance with human categorization the basic-level ontology scored best.

From these results we conclude that the most information will be gathered using the expert-based ontology when it is used by experts but the number of questions will be larger than when using other ontologies. The basic-level ontology is most compliant with human categorization but costs a lot of time to construct. The pragmatic-based ontology generates the least number of questions but the amount of information is the smallest of the three ontology.

An experiment was conducted to validate the framework with the exception of efficiency to construct the ontologies. Participants observed several videos and answered questions to determine an object on fire. It was shown that the basic-level ontology is most suitable for SAQG when the questions are posted to ordinary people. The expert-based ontology can be used when a more detailed concept is needed. This conclusion confirms the results when the framework to measure the compliance with human categorization is used to measure the suitability of the ontologies.

The answer to the general research question is that it is possible to automatically generate questions about a situation that are answered reliable by ordinary people. During the research though we encountered several subjects which gave rise to discussion such as what criteria for alignment with human categorization are available, whether other sources can be used to create a suitable ontology, how the information value or entropy of an ontology should be measured and whether microblogs, e.g. Twitter, are also suitable for information gathering.

Nederlandse samenvatting

Om de effecten van een crisis the verminderen, is het belangrijk zo snel mogelijk informatie te verzamelen. Gewone mensen die betrokken zijn bij een crisissituatie zijn tot op heden vaak genegeerd als bron van informatie. Tegenwoordig worden ze steeds meer beschouwd als de echte 'eerste reageerders'. In deze thesis wordt een raamwerk voorgesteld waarmee snel betrouwbare informatie van gewone mensen over een crisissituatie kan worden verzameld. Dit raamwerk is gerealiseerd in een toepassing: de Situation Awareness Question Generator (SAQG). SAQG genereert automatisch vragen van een ontologie en gebruikt de antwoorden om een situatie te bepalen. In dit raamwerk zijn vorm en inhoud van vragen bepaald door de specifieke ontology die gebruikt wordt, zijn er verschillende experimenten uitgevoerd om de ontologie die het meest geschikt is te karakteriseren.

De centrale vraag in dit proefschrift is als volgt: *Is het mogelijk automatisch vragen te genereren die beantwoord kunnen worden op zo een manier door gewone mensen dat de aard van een crisissituatie bepaald kan worden op een betrouwbare wijze?* Deze algemene vraag is onderverdeeld in de volgende specifieke onderzoeksvragen:

1. Wat is een geschikt raamwerk voor formalisatie van crisissituaties?
2. Gegeven een model om situaties te formaliseren, hoe kunnen we dit model gebruiken om relevante vragen te genereren?
3. Welke methoden zijn geschikt voor het ontwikkelen van een domein specifieke ontologie voor crisis situaties?
4. Welke methoden kunnen gebruikt worden om overeenkomst tussen domein specifieke ontologieën en natuurlijke categorisatie te meten?
5. Wat is de betrouwbaarheid van de antwoorden van gewone mensen op automatisch gegenereerde vragen over situaties?

Om situaties te formaliseren gebruiken we Situation Theory dat een rijkere structuur heeft dan andere gerelateerde formalismen en daardoor de mogelijkheid geeft tot een veel meer gedetailleerde representatie. In Situation Theory worden situaties gerepresenteerd door (samengestelde) infons. Door gebruik te maken van parameters kunnen infons abstraheren over verschillende mogelijke situaties in de echte wereld. Om een ontologie te maken hebben we Situation Theory vertaald naar Web Ontology Language (OWL). Hiervoor gebruikten we Situation Theory Ontology en hebben dat verbeterd om het meer in overeenstemming te laten zijn met prinicipes van kennisrepresentatie.

We hebben verschillende strategieën ontwikkeld om vragen automatisch te genereren. Karakteristieken van een ontologie bepalen welke strategie geschikt is. Wanneer de infons om een situatie te definiëren niet aan elkaar gerelateerd zijn, wordt een nogal moeizame strategie gebruikt. Wanneer een situatie wordt gedefinieerd door geparametriseerde infons is het afhankelijk van het type parameter welke strategie wordt gebruikt. Wanneer regels belangrijk zijn, wordt hiervan gebruik gemaakt zodat sommige vragen niet meer gesteld hoeven te worden. Voor een ontologie bestaande uit relaties gebaseerd op overerving wordt weer een andere strategie gebruikt. Deze strategieën zijn geïmplementeerd in SAQG, dat een algoritme gebruikt om een strategie te kiezen.

Voor een experiment dat moest uitwijzen welke rol mensen betrokken bij een crisissituatie willen innemen, ontwikkelden we een ontologie die verschillende rollen (Slachtoffer, Niet-Aktief, Waarnemer en Helper) representeert. Verschillende scenario's die een crisissituatie beschrijven, werden gepresenteerd aan deelnemers. Nadat ze deze scenario's hadden gelezen, beantwoordden de deelnemers vragen. De antwoorden werden gebruikt om de rol van elke deelnemer te bepalen. Het experiment toonde aan dat onze ontologie niet de vragen genereerde die leidden tot de rollen die we voor ogen hadden. We concludeerden dat onze ontologie niet overeenkwam met hoe gewone mensen mogelijke rollen tijdens een crisissituatie categoriseren.

Voor het ontwikkelen van een andere ontologie, dit keer van auto ongevallen, gebruikten we concepten genoemd door experts in relevante literatuur. Daarnaast betrokken we concepten uit twee databases die worden gebruikt door professionals op het gebied van auto ongevallen. Een experiment waarbij we SAQG gebruikten, toonde aan dat deze ontologie wél de vragen genereerde waarbij deelnemers betrouwbare antwoorden gaven, d.w.z. de antwoorden kwamen overeen met de werkelijke situatie.

Voor het karaktiseren van een ontologie die geschikt is voor het automatisch genereren van vragen, hebben we drie verschillende ontologieën ontwikkeld. Een pragmatisch-gebaseerde ontologie die we ontwikkelden uit een vocabulair, dat door de Amsterdamse brandweer wordt gebruikt om meldingen die binnenkomen bij 112 te categoriseren. Een expert-gebaseerde ontologie die we ontwikkelden met behulp van twee ontologieën gecreëerd door enerzijds kennisexperts en anderzijds experts op het gebied van kunst. Een basic-level ontologie werd geconstrueerd van concepten en hun attributen die we eerder gevraagd hadden aan gewone mensen. Door gebruik te maken van de attributen ontwikkelden we een ontologie met behulp van een algoritme dat gebaseerd was op de basic-level theorie van Rosch.

De bruikbaarheid van deze drie ontologieën werd gemeten in een raamwerk waarin we de ontologieën evalueerden op basis van vier aspecten: *a*) de ontologie moet een structuur hebben die geschikt is voor het automatisch genereren van vragen, *b*) de constructie van de ontologie moet efficiënt zijn, *c*) de ontologie moet compleet zijn en *d*) de ontologie moet overeenkomen met natuurlijke categorisatie. Voor dit raamwerk hebben we verschillende bestaande metrieken gebruikt zoals maximum padlengte, aantal concepten, entropie en het Ingve-Miller getal. Daarnaast hebben we een nieuwe metriek ontwikkeld die we 'semantische afstand validatie' hebben genoemd. Dit metriek vergelijkt de afstand van de concepten gemeten in padlengte met hoe deelnemers aan een experiment deze afstand evalueren.

Van deze metrieken hebben we geleerd dat de expert-gebaseerde ontologie een structuur heeft die het minst geschikt is voor het automatisch genereren van vragen, omdat het een groot aantal concepten, de grootste padlengte, het hoogste maximum aantal subconcepten en een hoge entropie heeft. De constructie van de basic-level ontologie was het minst efficiënt omdat het verkrijgen van de concepten en hun attributen veel werk vraagt. Het herontwikkelen van de vocabulaire om een pragmatisch-gebaseerde ontologie te verkrijgen, kostte meer tijd dan het toepassen van een algoritme zoals is gedaan bij het ontwikkelen van een expert-gebaseerde ontologie. De expert-gebaseerde ontologie is het meest compleet en de pragmatisch-gebaseerde ontologie het minst. Als het gaat om de overeenkomst met natuurlijke categorisatie scoort de basic-level ontologie het best.

Op basis van deze resultaten concludeerden we dat de meeste informatie wordt verkregen wanneer een expert-based ontologie wordt gebruikt wanneer deze wordt gebruikt door experts. Maar het aantal vragen zal groter zijn dan indien een andere ontologie wordt gebruikt. De basic-level ontologie komt het meest overeen met natuurlijk categorisatie, maar kost veel tijd om te ontwikkelen. De pragmatisch-gebaseerde ontologie genereert het minst aantal vragen, maar de hoeveelheid informatie die wordt verkregen is ook het minst.

Vervolgens hebben we een experiment uitgevoerd om dit raamwerk, m.u.v. de efficiency van het ontwikkelen, te valideren. Deelnemers werden diverse video's getoond en zij moesten vragen beantwoorden om de situatie te bepalen. In dit experiment werd aangetoond dat de basic-level ontologie het meest geschikt is voor SAQG wanneer de vragen worden gesteld aan gewone mensen. De expert-gebaseerde ontologie kan worden gebruikt wanneer meer gedetailleerde informatie nodig is. Deze conclusie bevestigt de resultaten van het raamwerk.

Het antwoord op de algemene onderzoeksvraag is dat het mogelijk is automatisch vragen te genereren over een situatie die betrouwbaar worden beantwoord door gewone mensen. Gedurende het onderzoek kwamen we diverse vragen tegen die aanleiding gaven tot discussie zoals welke criteria voor overeenkomst met natuurlijke categorisatie zijn beschikbaar, kunnen andere bronnen gebruikt worden het maken van een geschikte ontologie, hoe kan informatiewaarde of entropie van een ontologie gemeten worden en of microblogs zoals Twitter ook geschikt zijn voor het vergaren van informatie.

Bibliography

AATNed. AATNed, 2012.

N. H. T. S. Administration. Fatal Accident Reporting System. 2011. URL `http://www.nhtsa.gov/FARS`.

P. Adriaans. A critical analysis of Floridi's theory of semantic information. *Knowledge, Technology & Policy*, 23(1-2):41–56, 2010.

F. Baader and W. Nutt. Basic Description Logics. In F. Baader, D. Calvanese, D. L. McGuiness, D. Nardi, and P. F. Patel-Schneider, editors, *The Description Logic Handbook. Theory, Implementation and Applications*, pages 45–104. CUP, Cambridge, UK, 2007.

R. Baayen, P. R., and R. v. Rijn. The CELEX lexical data base on CD-ROM, 1993.

R. H. Baayen, R. Piepenbrock, and L. Gulikers. The CELEX Lexical Database (CD-ROM). Technical report, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

J. Barwise and J. Perry. *Situations and Attitudes.* MIT Press, Cambridge, MA, 1983.

S. Bechhofer and A. Miles. SKOS Simple Knowledge Organization System Reference. W3C recommendation, W3C (August 2009).

C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.

E. Bolotnikova, T. Gavrilova, and V. Gorovoy. To a method of evaluating ontologies. *Journal of Computer and Systems Sciences International*, 50(3): 448–461, 2011.

M. Bremer and D. Cohnitz. *Information and Information Flow.* Ontos Verlag, 2004.

C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. Data-driven Ontology Evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, 2004. European Language Resources Association.

P. Buitelaar, P. Cimiano, and B. Magnini. Ontology learning from text: an overview. In *Ontology learning from text: methods, evaluation and applications*, volume 123, pages 3–12. Ios Pr Inc, 2005.

J. Calmet and A. Daemi. From Entropy to Ontology. In *Cybernetics and Systems 2004-AT2AI-4: from Agent Theory to Agent Implementation*, 2004.

A. Campbell, S. Eisenman, N. Lane, E. Miluzzo, R. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G. Ahn. The rise of people-centric sensing. *Internet Computing, IEEE*, 12(4):12–21, 2008.

R. Carnap and Y. Bar-Hillel. An Outline of a Theory of Semantic Information. Technical report, MIT, October 1952.

J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson. Jena: implementing the semantic web recommendations. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 74–83, New York, NY, USA, 2004. ACM. ISBN 1-58113-912-8.

W. Chen, Y. Cai, H. Leung, and Q. Li. Generating ontologies with basic level concepts from folksonomies. *Procedia Computer Science*, 1(1):573–581, 2010.

M. Cho, C. Choi, W. Kim, J. Park, and P. Kim. Comparing Ontologies Using Entropy. In *Convergence Information Technology, 2007. International Conference on*, pages 873–876. IEEE, 2007.

L. Clarke. Panic: Myth or reality? *Contexts*, 1(3):21–26, Fall 2002.

E. U. Commission. Community database on Accidents on the Roads in Europe, 2011. URL `http://ec.europa.eu/transport/road_safety/specialist/-statistics/care_reports_graphics/index_en.htm`.

J. Corter and M. Gluck. Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, 111(2):291, 1992.

G. Coskun, M. Rothe, K. Teymourian, and A. Paschke. Applying community detection algorithms on ontologies for identifying concept groups. In *WoMO*, 2011.

S. D'Alfonso. *Towards a Framework for Semantic Information.* PhD thesis, University of Melbourne, Department of Philosophy, 2012.

M. Dean and G. Schreiber. OWL Reference. Technical report, W3C, 2004. URL `http://www.w3.org/TR`.

K. Devlin. *Logic and Information.* Cambridge University Press, Cambridge, UK, 1991.

I. Dillingham, J. Dykes, and J. Wood. Visual analytical approaches to evaluating uncertainty and bias in crowd sourced crisis information. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 273–274. IEEE, 2011.

P. Doran, V. Tamma, T. Payne, and I. Palmisano. An entropy inspired measure for evaluating ontology modularization. In *Proceedings of the fifth international conference on Knowledge capture*, pages 73–80. ACM, 2009.

M. Endsley. Theoretical Underpinnings of Situation Awareness: A Critical Review. In M. Endsley and D. Garland, editors, *Situation Awareness Analysis and Measurement*, pages 3–33. CRC Press, 2000.

D. Fensel. *Ontologies: a silver bullet for knowledge management and electronic commerce*. Springer Verlag, 2004.

L. Floridi. Outline of a Theory of Strongly Semantic Information. *Minds and Machines*, 14(2):197–221, 2004.

L. Floridi. *The philosophy of information*. Oxford Univ Pr, 2011.

M. Frixione and A. Lieto. Representing concepts in artificial systems: a clash of requirements. *Proc. HCP 2011*, pages 75–82, 2011.

B. Gaines and M. Shaw. Knowledge acquisition tools based on personal construct psychology. *The knowledge engineering review*, 8(1):49–85, 1993.

A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann. A theoretical framework for ontology evaluation and validation. In *SWAP*, volume 166, 2005.

J. García, F. J. García-Peñalvo, and R. Therón. A Survey on Ontology Metrics. In *Knowledge Management, Information Systems, E-Learning, and Sustainability Research*, pages 22–27. Springer, 2010.

A. Gatt and E. Reiter. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93, Athens, Greece, March 2009. Association for Computational Linguistics.

N. Goodman and R. Langhelm. Passive Disaster Reporting Through Mobile Social Networking Technology. In F. Fiedrich and B. Van de Walle, editors, *Proceedings of the 5th International ISCRAM Conference*. ISCRAM, May 2008.

M. Grobe. RDF, Jena, SparQL and the 'Semantic Web'. In *SIGUCCS '09: Proceedings of the 37th annual ACM SIGUCCS fall conference*, pages 131–138, New York, NY, USA, 2009. ACM.

T. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International journal of human computer studies*, 43(5):907–928, 1995.

N. Guarino. Toward a formal evaluation of ontology quality. *IEEE intelligent Systems*, 19(4):78–79, 2004.

N. Guarino and C. Welty. Evaluating ontological decisions with OntoClean. *Communications of the ACM*, 45(2):61–65, 2002.

D. Hall and J. Jordan. *Human-centered information fusion*. Artech House Publishers, 2010.

J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

A. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. In *Proceedings of the 6th International ISCRAM Conference*, May 2009.

A. Jaimes and S. Chang. A conceptual framework for indexing visual information at multiple levels. *IS&T/SPIE Internet Imaging*, 3964:2–15, 2000.

P. Kellman and P. Garrigan. Perceptual learning and human expertise. *Physics of life reviews*, 6(2):53–84, 2009.

G. Kelly. *Personal construct psychology*. Norton PressNY, 1955.

E. Keuleers, M. Brysbaert, and B. New. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, 42 (3):643–650, 2010.

H. Knublauch, M. Horridge, M. Musen, A. Rector, R. Stevens, N. Drummond, P. Lord, N. Noy, J. Seidenberg, and H. Wang. The Protégé OWL Experience. In *Fourth International Semantic Web Conference (ISWC2005)*, Galway, Ireland, 2005.

M. Kokar, C. Matheus, and K. Baclawski. Ontology-based situation awareness. *Information Fusion*, 10(1):83–98, 2009.

G. Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge Univ Press, 1987.

N. Lane, S. Eisenman, M. Musolesi, E. Miluzzo, and A. Campbell. Urban sensing systems: opportunistic or participatory? In *Proceedings of the 9th workshop on Mobile computing systems and applications*, pages 11–16. ACM, 2008.

H. Levesque, F. Pirri, and R. Reiter. Foundations for the Situation Calculus. *Computer and Information Science*, 3(18), 1998.

A. Lozano-Tello and A. Gómez-Pérez. Ontometric: A method to choose the appropriate ontology. *Journal of Database Management*, 2(15):1–18, 2004.

Y. Ma, B. Jin, and Y. Feng. Semantic oriented ontology cohesion metrics for ontology-based systems. *Journal of Systems and Software*, 83(1):143–152, 2010.

D. Macías-Galindo, W. Wong, L. Cavedon, and J. Thangarajah. Using a Lexical Dictionary and a Folksonomy to Automatically Construct Domain Ontologies. *AI 2011: Advances in Artificial Intelligence*, pages 638–647, 2011.

M. Maris and G. Pavlin. Distributed Perception Networks for Crisis Management. In *Proceedings of the 3rd International ISCRAM Conference (B.Van de Walle and M.Turoff, eds)*, 2006.

B. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

J. McCarthy. Situations, actions, and causal laws. Technical report, DTIC Document, 1963.

J. McCarthy and P. Hayes. *Some philosophical problems from the standpoint of artificial intelligence*. Stanford University USA, 1968.

G. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

M. Morris and G. Murphy. Converging operations on a basic level in event taxonomies. *Memory & Cognition*, 18(4):407–418, 1990.

G. Murphy. *The big book of concepts*. MIT Press, 2004.

G. Murphy and E. Smith. Basic-level superiority in picture categorization. *Journal of verbal learning and verbal behavior*, 21(1):1–20, 1982.

L. Palen, S. R. Hiltz, and S. B. Liu. Online forums supporting grassroots participation in emergency preparedness and response. *Commun. ACM*, 50 (3):54–58, 2007a.

L. Palen, S. Vieweg, J. Sutton, S. Liu, and A. Hughes. Crisis informatics: Studying crisis in a networked world.". In *Proceedings of the Third International Conference on E-Social Science*, 2007b.

L. Palen, K. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference*, page 8. British Computer Society, 2010.

I. Palmisano, V. Tamma, T. Payne, and P. Doran. Task oriented evaluation of module extraction techniques. *The Semantic Web-ISWC 2009*, pages 130–145, 2009.

A. Papasalouros, K. Kotis, and K. Kanaris. Automatic generation of multiple-choice questions from domain ontologies. *IADIS e-Learning*, 2008.

J. Payton and C. Julien. Integrating participatory sensing in application development practices. In *Proceedings of the FSE/SDP workshop on Future of software engineering research*, pages 277–282. ACM, 2010.

V. Pipek, L. Palen, and J. Landgren. Workshop summary: collaboration & crisis informatics (CCI'2012). In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 13–14. ACM, 2012.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1*, pages 448–453. Morgan Kaufmann Publishers Inc., 1995.

P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

A. Rifkin. Evidence for a basic level in event taxonomies. *Memory & Cognition*, 13(6):538, 1985.

E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976a.

E. Rosch, C. Simpson, and R. Miller. Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2 (4):491, 1976b.

M. Sabou and M. Fernandez. Ontology (network) evaluation. In *Ontology Engineering in a Networked World*, pages 193–212. Springer, 2012.

T. Scharrenbach, C. d Amato, N. Fanizzi, R. Grütter, B. Waldvogel, and A. Bernstein. Default Logics for Plausible Reasoning with Controversial Axioms. In *Proceedings of the 6th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2010)*, pages 105–108, November 2010.

G. Schreiber. Knowledge Engineering. In F. van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of knowledge representation*, Handbook of knowledge representation, chapter 25, pages 929–946. Elsevier, 2008.

G. Schreiber, H. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. Van de Velde, and B. Wielinga. *Knowledge engineering and management: the CommonKADS methodology*. MIT press, 1999.

J. Schultz and M. Liberman. Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of the DARPA broadcast news workshop*, pages 189–192. San Francisco: Morgan Kaufmann, 1999.

R. W. K. Sedgewick. *Algorithms*. Pearson Education, 2011.

Semiocast. http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US, June 2012.

C. Shannon. The Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

J. Smith and D. Kemler. Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General*, 113(1):137, 1984.

K. Starbird and L. Palen. Pass It On?: Retweeting in Mass Emergency. In *Proceedings of the 7th International ISCRAM Conference–Seattle*, volume 1, 2010.

A. Tapia, K. Bajpai, B. Jansen, J. Yen, and L. Giles. Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In *Proceedings of the 8th International ISCRAM Conference*, pages 1–10, 2011.

J. Taylor. *Linguistic categorization*. Oxford University Press, USA, 2003.

M. Teitsma, J. Sandberg, M. Maris, and B. Wielinga. Automatic question generation to determine roles during a crisis. In *SOTICS 2011, The First International Conference on Social Eco-Informatics*, pages 37–42, 2011a.

M. Teitsma, J. Sandberg, M. Maris, and B. Wielinga. Using an Ontology to Automatically Generate Questions for the Determination of Situations. In *Database and Expert Systems Applications*, pages 456–463. Springer, 2011b.

M. Teitsma, J. Sandberg, G. Schreiber, B. Wielinga, and W. R. van Hage. Engineering ontologies for question answering. *Applied Ontology*, 2014a.

M. Teitsma, J. Sandberg, B. Wielinga, and G. Schreiber. Validating Ontologies for Question Generation. In *Proceedings of the 26th Benelux Conference on Artificial Intelligence*, pages 113–120, November 6-7 2014b.

T. Terpstra, R. Stronkman, A. de Vries, and G. Paradies. Op weg naar een realtime twitter analyse tijdens crisis voor operationeel crisismanagement. *Ruimtelijke veiligheid en risicobeleid*, 3(6 & 7):57–65, 2012.

A. Tordai. *On Combining Alignment Techniques.* PhD thesis, VU University Amsterdam, 2012.

M. Turoff. Past and future emergency response information systems. *Commun. ACM*, 45(4):29–32, 2002.

B. Tversky and K. Hemenway. Categories of Environmental Scenes. *Cognitive psychology*, 15(1):121–149, 1983.

M. van Assem. *Converting and Integrating Vocabularies for the Semantic Web.* PhD thesis, VU University of Amsterdam, 2010.

J. van Benthem and M. Martinez. The stories of logic and information. In P. Adriaans and J. van Benthem, editors, *Handbook of the Philosophy of Information*, Handbook of the Philosophy of Information, pages 217–280. Elsevier, 2008.

J. van de Ven. The Combined Systems Project. Technical report, DECIS, Delft, The Netherlands, 2006. URL `http://combined.decis.nl`.

T. Van der Wouden, H. Hoekstra, M. Moortgat, B. Renmans, and I. Schuurman. Syntactic analysis in the spoken dutch corpus (CGN). In *Proceedings LREC*, 2002.

W. R. van Hage and D. Ceolin. The simple event model. In *Situation awareness with systems of systems*, pages 149–169. Springer, 2013.

W. R. Van Hage, V. Malaisé, R. Segers, L. Hollink, and G. Schreiber. Design and use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136, 2011.

G. van Heijst. *The role of ontologies in knowledge engineering.* PhD thesis, University of Amsterdam, 1995.

C. van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4): 365–373, December 1974.

S. Vieweg. Twitter communications in mass emergency: contributions to situational awareness. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 227–230. ACM, 2012.

P. Vossen, I. Maks, R. Segers, and H. van der Vliet. Integrating lexical units, synsets and ontology in the Cornetto Database. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2008.

D. Yue, S. Wang, and A. Zhao. Traffic Accidents Knowledge Management Based on Ontology. In *Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery - Volume 07*, FSKD '09, pages 447–449. IEEE Computer Society, 2009.

Y. Zhe, D. Zhang, and Y. Chuan. Evaluation metrics for ontology complexity and evolution analysis. In *e-Business Engineering, 2006. ICEBE'06. IEEE International Conference on*, pages 162–170. IEEE, 2006.

P. Zhou, Y. Zheng, and M. Li. How long to wait?: Predicting bus arrival time with mobile phone based participatory sensing. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 379–392. ACM, 2012.